IMPLEMENTATION OF REPUTATION BASED SELECTION OF WEB INFORMATION SOURCES

By

MOHSEN SOJOUDI 751336

HAMIDREZA SAEEDI 750835

Supervisor: Prof. Cinzia Cappiello

Master of Science in Management, Economics and Industrial Engineering

POLO REGIONALE DI COMO

Academic Year 2010/2011

IMPLEMENTATION OF REPUTATION BASED SELECTION OF WEB INFORMATION SOURCES

By

MOHSEN SOJOUDI 751336

HAMIDREZA SAEEDI 750835

Supervisor: Prof. Cinzia Cappiello

Master of Science in Management, Economics and Industrial Engineering

POLO REGIONALE DI COMO

Academic Year 2010/2011

ABSTRACT

The thesis introduces the reputation-based ranking of Web information sources and compares it with the Google's ranking. Moreover, it determines the relevance of the reputation metrics with respect to the Google's ranking algorithm. In the work, we focused on the blogs and forums since they allow users to share their opinions and insert their comments about the topics and assessing reputation for them is a crucial element.

The data quality literature defines reputation as a dimension of information quality that measures the trustworthiness and importance of an information source. Reputation is recognized as a multidimensional quality attribute. The variables that affect the overall reputation of an information source are related to the institutional clout of the source, to the relevance of the source in a given context, and to the general quality of the source's information content. A set of metrics measuring the reputation of Web information sources has been defined.

These metrics have been empirically assessed for the top 15 sources identified by Google as a response to ten queries in the tourism domain especially in New-York and London.

Then, we have compared Google's ranking with the reputation-based ranking for all the ten queries using different kinds of analysis. Results show that there is a difference (distance) between the Google's ranking and the ranking that is based on the reputation metrics. Moreover, the reputation metrics have different relevance to Google ranking algorithm since each ranking that is based along each of the reputation metrics has different distance values when comparing them with the Google's ranking.

At the next step the whole process is implemented as a web service. Our main focus

is in the areas of application implementation and enhancement, process optimization, interfaces and project management.

We have finally published our project over internet where you can access it on the following URL: www.ritrovatore.com

ACKNOWLEDGMENTS

This thesis arose in part out of years of research that has been done since we came to Politecnico di Milano. By that time, we have worked with a great number of people whose contribution in assorted ways to the research and the making of the thesis deserved special mention. It is a pleasure to convey our gratitude to them all in our humble acknowledgment.

In the first place we would like to record our gratitude to Prof. Cinzia Cappiello for her supervision, advice, and guidance from the very early stage of this research as well as giving us extraordinary experiences throughout the work. Above all and the most needed, she provided us unflinching encouragement and support in various ways. His truly scientist intuition has made her as a constant oasis of ideas and passions in science, which exceptionally inspire and enrich our growth as a student, a researcher and a scientist want to be. We are indebted to her more than she knows.

Where would we be without our family? Our parents deserve special mention for their inseparable support and prayers. Our Fathers, in the first place are the persons who put the fundament our learning character, showing us the joy of intellectual pursuit ever since we were a child. Our Mothers are the one who sincerely raised us with her caring and gently love.

Finally, we would like to thank everybody who was important to the successful realization of thesis, as well as expressing our apology that I could not mention personally one by one.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 11

Information is everywhere 11
Search engine 12

II. STATE OF THE ART 15

The most important Google ranking factors 24
A graphical concept of reputation 29
Reputation 30
Four dimensions to dependability 30
Traffic 31
Breadth of contributions 33
Relevance 34
Liveliness 36
Reputation system is everywhere 37
Reputation system affects our lives 37
Who's using this system? 38
Why we should use reputation system? 39
Why we need to design web reputation system? 39

III. IMPLEMENTAION 41

Main part 42
Query part 45
Retrieve part 49
Class part 52
Class Google 53
Class Yahoo 53
Class Alexa 54
Class Main 54
Functions part 55
Screen pop-up 56
Global rank 56
Country rank 57

Daily visit 57
Time on site 58
Average load time 58
Page speed score 59
Facebook mentions 59
Twitter mentions 60
Alexa back-link 60
Yahoo back-link 60
Bounce rate 61
Page view per user 61
Search visit 61
Re-ranking part 62
What is the AHP and how did we apply AHP to our model? 63
AHP steps for applying 64
AHP value 71

IV. CONCLUSION 75

V. USER MANUAL 77

Getting started 77
Browsing information's 78
Domains 78
Sub-domains 79
Cities 79
Query result 80
Google rank 80
Web reputation based rank 83

REFERENCES 84

LIST OF TABLES

Table Page

2.1 36
2.2 37
3.1 56
3.2 57
3.3 57
3.4 58
3.5 59
3.6 59
3.7 59
3.8 60
3.9 60
3.10 61
3.11 61

LIST OF FIGURES

Figure Page

2.1 20
2.2 29
3 39
3.1 44
3.2 47
3.3 48
3.4 51
3.5 54
3.6 64
3.7 65
3.8 66
3.9 67
3.10 68
3.11 68
3.12 69
3.13 70

1. INTRODUCTION

- ***Information is everywhere***

Finding information is difficult day by day due to the explosion of content that has gathered from different computer networking and data bases. Broadly distributed information makes it harder to issue a single request and get optimal result. Even in traditional applications were several and lot of database to be searched. So when a user has a request it must select which databases are more relevant, issue a query to any of those database and at the end review all of the appear result and find the top matching document as a request. The same similar problem arises in the case of sensor networks.

Some sensor networks were interested to make some of collected data publicity available on World Wide Web. So in this case, even large general search engines fails to analysis the growing content. Thus there is new trend toward specialized searched engines that they must find new solution to solve this mass problem. Also in addition, beyond the issue of scalability, a lot of website provides dynamic content that at that time those web crawler of search engine cannot access those information by hyperlink. For example consider a keyword search page that provide a specific result for user; as all of us know the web crawler only consider the text on the search filed that a user defined but it has no hyperlink for this document to redirect user to this document. So we reach to the non-crawlable content that becomes part of what is commonly today known as HIDDEN WEB. Similar arise in the sensor network.

At those day a lot of different approach were utilized, on approach was to dispatch the query to each information source that is likely to have requested documents and merge the final result before displaying to the user and they call this approach as **meta-searching**. Another was search engines.

- ***Search engine***

Web search has been an important tool in our life. Today, there are billions of web pages accessible on the Internet. These web pages are highly diversified in terms of content, format and quality. It is the basic challenge for search engines to rank web pages for a given query to find the most relevant one from such a huge amount of diversified data (Chen 2008). Web browsing most often starts from search engines and moves along a chain of links originating in the top search results.

A search engine is computer software that is continually modified to avail of the latest technologies in order to provide improved search results. Each search engine does the same function of collecting, organizing, indexing and serving results in its own unique way, thus employing various algorithms and techniques, which are their trade secrets. In short, the functions of a search engine can be categorized into the following broad areas: First, crawl the Web and locate all Web pages. Second, index the data. Third, rate the importance of each page in the database, so that when a user does a search and the subset of pages in the database with the desired information has been found, the more important pages can be presented first (Gupta, S.; Jindal, A. 2008).

Search engines are general purpose and implement proprietary ranking algorithms which, although efficient and commonly effective, do not always meet users' expectations. Users are often dissatisfied with the ability of search engines to identify the best information sources within a given domain or for a given purpose. It is common experience how the identification of relevant information on a specific issue through Web browsing requires several iterations and interesting sources may surface as a result of relatively long search sessions.

In (Jiang et al., 2008), empirical evidence is provided indicating that there is a quite large probability (about 63%) of a relevant document being found within a 1-120 rank range. In addition to that, the study found that the most relevant document in

substantially more than 65% of the cases, not even the top 300 ranked documents are expected to suffice.
Also in (Jiang et al.,2008), the rank range of documents users view (in the result list for a single query) has been widely studied, showing that users tend to look only at the first ten results and most of the users (percentages close to 80%) tend to not look deeper than two result pages.

The ranking algorithms used by search engines are authority based, i.e. they tie a site's ranking to the number of incoming Web links. This thesis explores the possibility of adjusting the ranking provided by search engines by assessing the reputation of Web information sources and by using the reputation metrics as a basis of the ranking thus improving the ranking process since users can find the relevant web information sources they are seeking for in less time because they will be ranked in the first positions in the retrieved list according to the required query inserted in the search engine and these reputation metrics take in to account the effective interaction between the users and the web information sources. Reputation is the opinion (more technically, a social evaluation) of the group of entities toward a person, a group of people, or an organization on a certain criterion. It is an important factor in many fields, such as education, business, online communities or social status. Reputation can be considered as a component of the identity as defined by others.

The data quality literature defines reputation as a dimension of information quality that measures the trustworthiness and importance of an information source. Reputation is recognized as a multi-dimensional quality attribute. In the data quality field, the concept of reputation is the result of the assessment of several properties of information sources, including correctness, completeness, timeliness, dependability, and consistency.

The variables that affect the overall reputation of an information source are related to the institutional clout of the source, to the relevance of the source in a given

context, and to the general quality of the source's information content. To the current state of the art, the literature lacks evidence demonstrating the importance of the concept of reputation in improving the ranking provided by search engines. It also lacks an operationalization of the concept of reputation allowing the assessment of Web information sources.

Our thesis focuses on information source selection based on reputation system where the honestly and important of data source and also the limitation and restriction of available high-quality data source need to be taken into consideration.

Chapter 1 will describe the main definition and concept of web reputation system. Chapter 2 will describe the design of a general web reputation system and discuss the issues that arise with the introduction of high-quality data source constraint and different available metric that must be consider for defining the quality of information data source. Chapter 3 will present implementation of our approach for information source selection based on web reputation. And finally conclusion and future research direction are presented in chapter 4.

2. STATE OF THE ART

Web 2.0 technologies which are commonly associated with web applications that facilitate interactive information sharing and collaboration on the World Wide Web enable an active role of users, who can create and share their contents very easily. A Web 2.0 site gives its users the free choice to interact or collaborate with each other in a social media dialogue in a virtual community, in contrast to websites where users (consumer) are limited to the passive viewing of content that was created for them. This mass of information includes opinions about a variety of key interest topics (e.g., products, brands, services, or any subject of interest for users) and represents a new and invaluable source of marketing information.

Web 2.0 technologies also allow people to express their opinions, and distribute them through several means (e.g., forums, blog posts, social networks), thus increasing the amount of information on the Web. Nowadays, companies find the Web as an important resource for checking customers' appreciation for their products, services and even to understand their brands' reputation, since it is well known that online reviews can have a negative impact on sales and Weblog mentions are highly correlated with sales.

Search engines are the key to finding specific information on the vast expanse of the World Wide Web. Without sophisticated search engines, it would be virtually impossible to locate anything on the Web without knowing a specific URL. When people use the term search engine (A program that searches documents for specified keywords and returns a list of the documents where the keywords were found) in relation to the Web, they are usually referring to the actual search forms that searches through databases of HTML documents, initially gathered by a robot (A program that runs automatically without human intervention).

There are basically three types of search engines: Those that are powered by robots (Called crawlers; ants or spiders) and those that are powered by human submissions and those that are a hybrid of the two;

Crawler-based search engines are those that use automated software agents (called Crawlers) that visit a Web site, read the information on the actual site, read the site's Meta tags (A special HTML tag that provides information about a Web page) and also follow the links that the site connects to performing indexing on all linked Web sites as well. The crawler returns all that information back to a central depository, where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed. The frequency with which this happens is determined by the administrators of the search engine.

Human-powered search engines rely on humans to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index. In both cases, when you query a search engine to locate information, you're actually searching through the index that the search engine has created — you are not actually searching the Web. These indices are giant databases of information that is collected and stored and subsequently searched. This explains why sometimes a search on a commercial search engine, such as Yahoo! Or Google, will return results that are, in fact, dead links. Since the search results are based on the index, if the index hasn't been updated since a Web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated.

The classic document-ranking technique involved viewing the text on a website and determining its value to a search query by using a set of so-called "on-page" parameters. A simple text-only information retrieval system produces poor search results. In the past, several text-only search engines relied upon on-page ranking factors.

One of the early web crawlers was Wandex, created in 1993 at MIT by Matthew Gray. WebCrawler, released in 1994, is considered the first web crawler to look at the entire text of a web document.

When ranking a document, the early companies (and most that followed) focused on what are now called "on-page factors"—parameters a webpage author can control directly. These parameters are of little use in generating relevant search results.

If we were to write a crude ranking algorithm we could create combinations of HTML parameters appearing on a webpage to generate ranking factors. By using on-page HTML parameters, a simple ranking algorithm could generate a list of relevant documents to a given search query. This approach has the built-in assumption that the authors of the WebPages we are indexing are honest about the content they are authoring. An algorithm is simply a set of instructions, usually mathematical, used to calculate a certain parameter and perform some type of data processing. It is the search engine developer's job to generate a set of highly relevant documents for any search query, using the available parameters on the web. The task is challenging because the available parameters usable by the algorithm are not necessarily the same as the ones web users see when deciding if a webpage is relevant to their search.

Looking at the available parameters in an HTML document, one can derive a list of potential on-page variables for ranking web documents. For example, in the early 1990s, a search engine called Veronica used the index results from a program called Gopher to look at webpage titles and URLs to determine the topic and relevance of a webpage.

Because the document's author can easily manipulate the title of a web document and it's URL, a good ranking algorithm would require either more variables or rely on factors a webpage author cannot control directly. Using more variables in a

ranking algorithm naturally makes the manipulation of its search results more difficult.

The following represents the potential on-page factors:

1. Description Meta tag (A special HTML tag that provides information about a Web page).
2. A website's URL.
3. The title of a website.
4. Keyword Meta tags.
5. Density of a given keyword on a document.
6. Proximity of keywords defines how close keywords are in relation to each other.
7. Prominence of keywords defines where the keywords are on the HTML page. For example, a keyword with high prominence would be at the top of an HTML document.
8. Keywords using HTML bold and/or italics.
9. Overall size of a page.
10. Total number of pages within the website.
11. Number of outbound links.
12. Use of quotes text keywords.
13. Using underscores on text keywords.
14. The uniqueness of the content on your page relative to the other content on the web.
15. Content "freshness." When was content last updated? Has it changed since the last time it was crawled?
16. Spelling and grammar.

A one-dimensional search algorithm might calculate the density of a keyword on a page, and use that keyword density as a measure of relevance. This type of search can quickly lead to text manipulation if the web authors are aware that they need

simply to change the keyword density of their web document to indicate to a search engine what their document is about.

Using only the on-page factors, webspam will be difficult to stop because the website optimizer can still control the parameters the search algorithm is using to determine ranking.

To this extent, off-page factors were introduced. These factors are difficult for the webpage optimizer to control. Off-page metrics are more desirable in any ranking algorithm because they allow the search algorithm to determine which pages appear in search queries, rather than by webpage optimizers manipulating WebPages.

The following represent the potential off-page factors:

1. Number of websites linking back to a website.
2. The page-rank of a website.
3. The number and quality of directories a page is listed in. For example DMOZ or Yahoo;
4. How long a URL has been registered?
5. When a registered domain name will expire.
6. When the search engine spider last crawled the URL.
7. How many pages of the website were crawled (crawl depth).
8. How fast the pages can be crawled (crawl rate).

One reason for moving to metrics like those is that they are less obvious to the website optimizer. Major search engines like Google and Yahoo! Have a majority of the world's search queries at their disposal. These search engines also have access to statistical data for how authoritative WebPages have evolved over time. Armed with this type of information, search engines can develop algorithms that can detect unnatural webpage behavior.

The life span of a Google query normally lasts less than half a second, yet involves a number of different steps that must be completed before results can be delivered to a person seeking information.

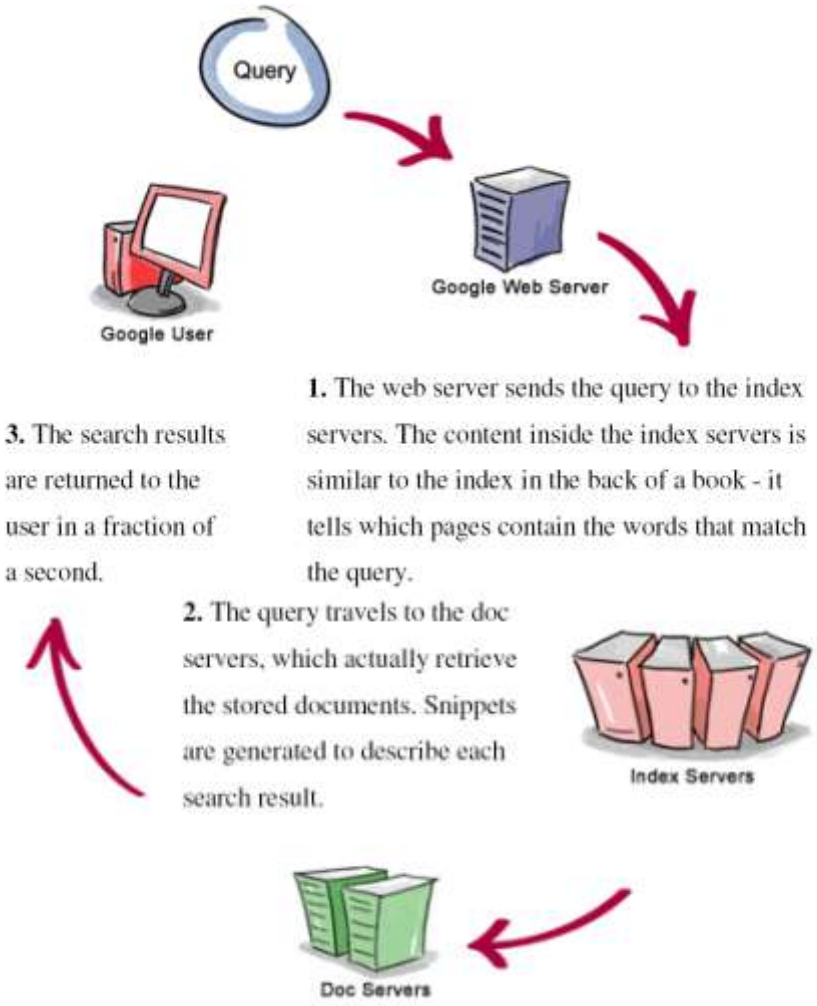

Query
Google User
Google Web Server
1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query.
2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.
Index Servers
Doc Servers
3. The search results are returned to the user in a fraction of a second.

Figure 1 Life span of a Google query

Figure 2.1

So why will the same search on different search engines produce different results? Part of the answer to that question is because not all indices are going to be exactly the same. It depends on what the spiders find or what the humans submitted. But more important, not every search engine uses the same algorithm to search through the indices. The algorithm is what the search engines use to determine the relevance of the information in the index to what the user is searching for.

Also some search engines index more web pages than others. Some search engines also index web pages more often than others. The result is that no search engine has the exact same collection of web pages to search through. That naturally produces differences, when comparing their results.

Search engines may also penalize pages or exclude them from the index, if they detect search engine "spamming." An example is when a word is repeated hundreds of times on a page, to increase the frequency and propel the page higher in the listings. Search engines watch for common spamming methods in a variety of ways, including following up on complaints from their users.

One of the main rules in a ranking algorithm used by search engines involves the location and frequency of keywords on a web page. Call it the location/frequency method, for short.

Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant than others to the topic. Search engines will also check to see if the search keywords appear near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning.

Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in

a web page. Those with a higher frequency are often deemed more relevant than other web pages.

Crawler-based search engines have plenty of experience now with webmasters who constantly rewrite their web pages in an attempt to gain better rankings. Because of this, all major search engines now also make use of "off the page" ranking criteria.

Off the page factors are those that a webmasters cannot easily influence. Chief among these is link analysis. By analyzing how pages link to each other, a search engine can both determine what a page is about and whether that page is deemed to be "important" and thus deserving of a ranking boost. In addition, sophisticated techniques are used to screen out attempts by webmasters to build "artificial" links designed to boost their rankings.

Another off the page factor is click through measurement. In short, this means that a search engine may watch what results someone selects for a particular search, and then eventually drop high-ranking pages that aren't attracting clicks, while promoting lower ranking pages that do pull in visitors.

Other algorithms that are used by search engines are the Page-Rank and the Hyperlink-Induced Topic Search (HITS). Page-Rank is a link analysis algorithm, named after Larry Page, used by the Google Internet search engine. Link analysis is a subset of network analysis, exploring associations between objects. It provides the crucial relationships and associations between very many objects of different types that are not apparent from isolated pieces of information.

In short Page-Rank is a “vote”, by all the other pages on the Web, about how important a page is. A link to a page counts as a vote of support. If there’s no link there’s no support.

Page-Rank algorithm is in fact elegantly simple and is calculated as follows:

- **PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn))**

Where PR (A) is the Page-Rank of a page A

PR (T1) is the Page-Rank of a page T1

C (T1) is the number of outgoing links from the page T1

d is a damping factor in the range 0 < d < 1, usually set to 0.85

The Page-Rank of a web page is therefore calculated as a sum of the Page-Rank s of all pages linking to it (*its incoming links*), divided by the number of links on each of those pages (*its outgoing links*).

Page-Rank can affect the position of your page on Google in two ways:

- *The number of incoming links.* Obviously the more of these the better
- *The number of outgoing links* on the page which points at your page. The fewer of these the better. This is interesting: it means given two pages of equal Page-Rank linking to you, one with 5 outgoing links and the other with 10, you will get twice the increase in Page-Rank from the page with only 5 outgoing links.

Hyperlink-Induced Topic Search (HITS) (also known as Hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

The most important Google ranking factors are:

1. **A**ge of Domain: Age of URL is very important. If you just bought your domain a few weeks or even months ago you have a long road ahead of you. The reality is the age of your website helps build trust.

2. **D**omain Hosting: Where is your site hosted? Find out through your hosting company what continent or country your site is hosted in. This can often times play a large role in search rankings. Always use a reputable hosting company. Never use the cheapest hosting. The reality is, if you cannot afford hosting you should re-consider the business.

3. **Y**our Neighbors: Make sure that your neighbors on your server are not classified as spam.

4. **U**RL Structure: Make sure your URL structures are very clean. There should not be any random strings of characters at the end of your URL's.

5. **C**ontent: Content is very important. To start make sure you have text on all your important pages, then make sure it is good text consisting of your targeted keywords spread throughout naturally. Simply put, ALWAYS write your content for humans, your website visitors first and NEVER write content for the solo purpose to achieve Google search engine rankings.

6. **I**nternal Link Structure: Make sure your inner pages are linked correctly. Visitors should have easy made pathways connecting to your other pages from every page of your website. Essentially make sure the site is clean, easy to use and interlinked to help the user experience.

7. **T**rust: Do you at least have a mailing address listed on your website? You should if you don't. Google likes to see trust factors on websites so anything you can add that could help build trust for your audience will benefit your rankings. Make it easy for people to do business with you, it all starts with establishing trust and that starts with contact information on your website.

8. **K**eywords: Make sure your website is optimized using your keywords. Remember to naturally optimize your website based on the content of each page of your website.

9. **B**ounce Rate: Although bounce rate might not seem important if Google sees that nobody hangs out on your website for more than a few seconds before they leave this could be a ranking problem over time. Make changes to get visitors engaged with your website. Simple things, like video, newsletter sign up, call to actions, etc will help improve your bounce rate over time.

10. **O**utbound links: Make sure the websites that you link to are 100% relevant to your business and industry. If you sell animals toys but you are linking to a site that sells shoes that is not very relevant

and over time could really impact your rankings. Bottom line is if it makes sense to link to another site, then do so, but remember you could be sending your visitors away from your site.

11. **I**nbound Links: The key here don't buy or exchange links. Market and promote your business online to build visitors to your website over time. If you do, then the relevant links will follow!

12. **P**age views: One of the most fundamental starting points for measuring the performance of your content is looking at how many views or hits it receives. The pages that are viewed or landed on most often on your Website can give you clues into what people are searching for and what information they find valuable. On the other hand, tracking views can also tell you what pages are under-performing.

13. **C**omments/feedback: Comments are a great way to track response to your content, and they offer you insight into what your community is interested in learning more about, what questions they have, and where they believe you, as an organization, can and do fill in the knowledge gaps. If you find that a specific topic you've blogged about gets a significant number of questions in the comments, those questions can directly impact future content you produce surrounding that topic, and tell you what's missing in the greater industry picture.

14. **L**ength of visit: The length of time someone spends (or doesn't spend) with your content can be a strong indicator of your content's success or shortcomings. That understanding can be

identified as an actual visit, instead of a bounce; clicks through to other pieces of your Website can also add to the time someone spends on it and identify their interest in your service or product.

15. **S**hares: If your content is hitting home with your community, there's a good chance individuals are sharing it with their colleagues, teams, and networks. For instance, say you received a great email newsletter today from a vendor that you know a decent portion of your Twitter network would benefit from seeing.

The proposed approach suggests that ranking should be based on the reputation metrics (a multi-dimensional ranking) which take into account the effective interaction between the users and the information sources. It is common experience how the identification of relevant information on a specific issue through Web browsing requires several iterations and interesting sources may surface as a result of relatively long search sessions. Empirical evidence is provided indicating that there is a quite large probability (about 63%) of a relevant document being found within a 1-120 rank range. In addition to that, the study found that the most relevant document in substantially more than 65% of the cases, not even the top 300 ranked documents are expected to suffice (Jiang et al., 2008). By adopting the proposed approach, it will allow users to find the information that they are seeking for in less time since the most relevant websites will be ranked in the first positions because the reputation based ranking will take in to consideration the effective interaction between the users and the Web information sources.

The selection of sources providing dependable information has been scarcely based on the definition of methods for assessing Data Quality (DQ). Data are of high quality "if they are fit for their intended uses in operations, decision making and planning". Alternatively, the data are deemed of high quality if they correctly represent the real world construct to which they refer. Furthermore, apart from these definitions, as

data volume increases, the question of internal consistency within data becomes paramount, regardless of fitness for use for any external purpose In the DQ field, the concept of reputation is the result of the assessment of several properties of information sources, including correctness, completeness, timeliness, dependability, and consistency.

Reputation is recognized as a multidimensional quality attribute. Data accuracy refers to the degree with which data correctly represents the “real-life” objects they are intended to model. In many cases, accuracy is measured by how the values agree with an identified source of correct information (such as reference data). There are different sources of correct information: a database of record, a similar corroborative set of data values from another table, dynamically computed values, or perhaps the result of a manual process. Completeness of data is the extent to which the expected attributes of data are provided. Data completeness refers to an indication of whether or not all the data necessary to meet the current and future business information demand are available in the data resource.

Data Completeness is the 'expected completeness'. It is possible that data is not available, but it is still considered completed, as it meets the expectations of the user. Every data requirement has 'mandatory' and 'optional' aspects. For example customer's mailing address is mandatory and it is available and because customer’s office address is optional, it is OK if it is not available.

Timeliness refers to the time expectation for accessibility and availability of information. It can be measured as the time between when information is expected and when it is readily available for use. Timeliness is affected by three factors: How fast the information system state is updated after the real-world system changes (system currency); the rate of change of the real-world system (volatility); and the time the data is actually used. While the first aspect is affected by the design of the information system, the second and third are not subject to any design decision.

A graphical concept of reputation

The phrase reputation system describes a wide array of practices, technologies, and use-interface elements. You'll notice that reputation system compute many different reputation values that turn out to possess a single common element: the reputation statement.

In practice, most input to a reputation model is either already in the form of reputation statements or quickly transformed into them for easy processing.

The reputation statement is like an atom in that it too has constituent particles: a source, a claim, and a target (figure). The exact characteristics (type and value) of each particle determine what type of element it its and its use in your application.

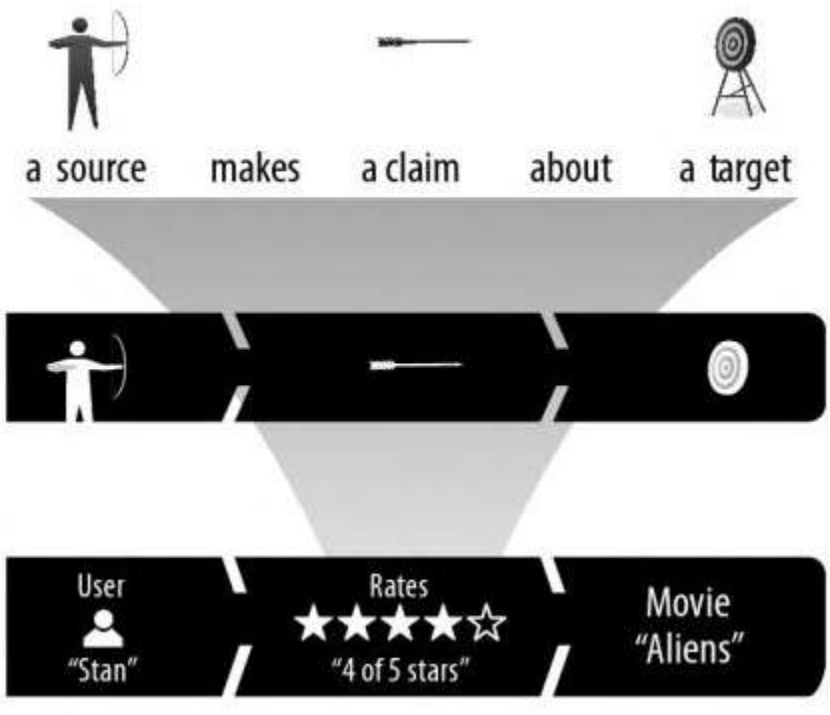


Figure 2.2

There are four dimensions to dependability:

1. **A**vailability: the availability of a system is the probability that it will be up and running and able to deliver useful services at any given time.
2. **R**eliability: the reliability of a system is the probability, over a given period of time, that the system will correctly deliver services as expected by the user.
3. **S**afety
4. **S**ecurity: the security of a system is a judgment of how likely it is that the system can resist accidental or deliberate intrusion.

Consistency of Data means that data across the enterprise should be in synch with each other. Consistency refers to data values in one data set being consistent with values in another data set. A strict definition of consistency specifies that two data values drawn from separate data sets must not conflict with each other in the proposed approach; reputation metrics were identified that are based on the data quality dimensions.

These metrics have been empirically assessed for the top 15 sources identified by Google as a response to ten queries in the tourism domain in Milano and London. Then, we have conducted several analyses to compare Google's ranking and the ranking that is based on reputation metrics for all the queries in order to assess the distance between the two different ranking algorithms and to measure the relevance of the reputation metrics with respect to the Google's ranking algorithm.

Two methods are used to measure the distance between the rankings which are the Spearman's Footrule distance and Kendall tau distance. To the current state of the art, the literature lacks evidence demonstrating the importance of the concept of reputation in improving the ranking provided by search engines. It also lacks an

operationalization of the concept of reputation allowing the assessment of Web information sources.

The proposed approach is based on the reputation-based selection of relevant and reliable Web information sources. Common experiences of users searching the Web reveal how the identification of relevant information on a specific issue through Web browsing requires several iterations and interesting sources may surface as a result of relatively long search processes. In (Jiang et al. 2008), empirical evidence reveals that there is a quite large probability (about 63%) of a relevant document being found within a 1-120 rank range, but also that, in more than 65% of the cases, not even the top 300 ranked documents are expected to satisfy the user request. Also in (Jiang et al.,2008), the rank range of documents users view (in the result list for a single query) has been widely studied, showing that users tend to look only at the first ten results and most of the users (percentages close to 80%) tend to not look deeper than two result pages.

The approach tries to overcome the previous problem by proposing the adoption of typical data quality dimensions to assess the reputation of information sources; this in turn allows ensuring a major quality of the retrieved information. The operationalization of reputation draws from the data quality literature. In data quality literature, Accuracy, Completeness, and Time represent the fundamental data quality dimensions in most contexts. Interpretability, Authority, and Dependability are suggested as additional dimensions that should be considered.
The four aspects that should be evaluated to assess the reputation of blogs and forums (two important forms of Web resources providing large amounts of users' opinions) are:

1) **T**raffic: overall volume of information produced and exchanged in a given time frame. Web traffic is the amount of data sent and received by visitors to a web site. It is a large portion of Internet traffic. This is determined by the number of

visitors and the number of pages they visit. Sites monitor the incoming and outgoing traffic to see which parts or pages of their site are popular and if there are any apparent trends, such as one specific page being viewed mostly by people in a particular country.

There are many indicators to measure traffic to your website, here are some of them:

a. ***Traffic rank:***

The Alexa traffic rank is calculated using a combination of average daily reach and page views. What is "reach"? We'll have more to say about this in an upcoming post, but basically it measures how many people are visiting a site, expressed as a fraction of the global Internet population. For example, if you click on the "Reach" link below the "Traffic Stats" tab, you can see that Google's reach is currently around 33%, meaning that about one in three Internet users visit google.com on a typical day. (Alexa.com)

b. ***Global rank:***

An Alexa global ranking is an indicator used to gauge site performance and appears to be popularity based which is achieved by users loading a search tool on their website.

c. ***Country rank:***

Country Rank" that indicates how much traffic it gets per country. If the Alexa rank is 100,000 or less, you can be confident that you will get some traffic from it.

d. Rank by country:

The rank by country is calculated using a combination of average daily visitors and page views from users from that country over the past 3 months.

e. Daily page views:

Page views measure the number of pages viewed by Alexa Toolbar users

f. Daily visitor:

i. *Number of users that visit website each day*

g. Global reach:

i. *Reach measures the number of users. Reach is typically expressed as the percentage of all Internet users who visit a given site. So, for example, if a site like yahoo.com has a reach of 28%, this means that if you took random samples of one million Internet users, you would on average find that 280,000 of them visit yahoo.com.*

h. Average load time:

*i. This is the **average load time** of the pages from your website or blog.*

2) **B**readth of contributions: overall range of issues on which the source can provide information. The more the issues that a source can cover, the more comprehensive the source is. Here are some of the ways that can be used to measure the contributions of the users to your forum or blog:

a. ***Face book mention:***

You can mention Friends, Pages, events, groups or apps in Face book. We count the number of mention links over Face book and it's a very good content performance indicator to understand the breath of contribution of website over social network.

b. ***Twitter mention:***

You can mention Friends, Pages, events, groups or apps in Twitter. We count the number of mention links over Twitter.

3) **R**elevance: degree of specialization of the source in a given domain (e.g. tourism). It also means the distance between the content of a Web resource and a user's query. The two most important features of your web page are the URL and the TITLE tag. These are heavily weighted for relevance by the search engines. Your URL should reflect the name of your business or site, or type of business, service or content on your site. The key is to keep the URL relevant to what you are doing. The other most important thing to do is to customize the TITLE tag of your html to reflect the content on your page or the message you want to convey. Again, it needs to be relevant to the content on the same page. Many novice website completely overlook the TITLE tag and miss out on higher rankings simply because they do not include a relevant description of the page in the TITLE header tag. To this extent, here are the most important guidelines that should be taken into consideration in the design and the content of a source.

a. ***Bounce rate:***

Bounce rate is leaving out the first page without going into deeper. It is the rate at which the visitors leave your website without really examining what it is about or not completing a particular activity or transaction. 80% bounce rate may either

refer to the number or proportion of visitors who visited your site and left (without doing anything). What is your website about? Maybe you can do a little improvement so it can stir the interest of your visitors.

b. *Time on site:*

Time on site is the length of visit on your website. A high time on site may indicate your visitors may be interacting extensively with your site. However, high time on site can be misleading:

Your visitors may have a hard time looking for what they want your visitor's leaves their browser windows open when they are not actually viewing or using your website

c. *Search visit keyword:*

Identifying these keywords in your research and targeting them on your landing pages will help you cherry pick the best traffic from the search engines–traffic that converts well

d. *Yahoo inbound link:*

Back-links are incoming links to a website or web page. Inbound links were originally important (prior to the emergence of search engines) as a primary means of web navigation; today their significance lies in search engine optimization (SEO). The number of back-links is one indication of the popularity or importance of that website or page (for example, this is used by Google to determine the Page Rank of a webpage). Outside of SEO, the back-links of a webpage may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page. (alexa.com)

4) **L**iveliness: responsiveness to new issues or events. You can measure the liveliness of your website in several ways such as

   a. ***Number of daily page views per daily visitor(Page view/ user)***

      The page views per user numbers are the average numbers of unique pages viewed per user per day by the users visiting the site.

From the above four variables (Traffic, Breadth of contributions, Relevance and Liveliness) and the data quality dimensions, we have identified the reputation metrics that should be measured to assess the reputation of a Web information source.

Table 2.1 summarizes the reputation metrics that were identified for the variables above (table columns) along the different data quality dimensions (table rows). As a general observation, the choice of metrics has been driven by feasibility considerations.

In particular, only quantitative and measurable metrics were defined. The data source on which metrics are computed is reported in parentheses. “Crawling” means either manual inspection or automated crawling depending on the site. Some metrics are also derived from data published by Alexa (www.alexa.com), a well-known service publishing traffic metrics for a number of Internet sites. It is worth noting that not all data quality dimensions apply to all variables (not applicable, N/A in Table 2.1).

	Traffic	*Breadth of contributions*	*Relevance*	*Liveliness*
Accuracy	n.a.	average number of comments to selected post (crawling)	Centrality, i.e., number of covered topics (crawling)	n.a.
Completeness	n.a.	number of open discussions (crawling)	number of open discussions compared to largest Web blog/forum (crawling)	number of comments per user (crawling)
Time	traffic rank (www.alexa.com)	age of source (crawling)	n.a.	average number of new opened discussions per day (www.alexa.com)
Interpretability	n.a.	average number of distinct tags per post (crawling)	n.a.	n.a.
Authority	- daily visitors (www.alexa.com) - daily page views (www.alexa.com) - average time spent on site (www.alexa.com)	n.a.	- number of inbound links (www.alexa.com) - number of feed subscriptions (Feedburner tool)	number of daily page views per daily visitor (www.alexa.com)
Dependability	n.a.	number of comments per discussion (crawling)	bounce rate (www.alexa.com)	average number of comments per discussion per day (crawling)

Table 2.1

Reputation system is everywhere

You have challenge with reputation system every day, even when you don't realize it. You can use reputation for life efficiency because reputation helps you make better judgment for better information. Reputation is very important over internet because of extendibility of the pages to be sort base on your needs and attention. Without reputation system for things like search ranking, rating, and review and also spam filter the web get unusable long ago.

Our project tries to clarify the important of the concept of web reputation in selecting web information sources and implementation of following concept over net. These tools will give better view to analyzers to understand the best way of choosing information source and its roles in raking result.

Reputation system affects our lives

We use reputation every day to make better decision about our daily normal or critical events. Now a day reputation system can evaluate your performance and

your creation. This effect is also true for the groups that you are member of it like society, work or others. They all have aggregate point that reflects you as well as the others. The group reputation systems are difficult most of the time and a hard to perceive and most of the time harder to change.

Who's using this system?

Some of the best-known consumer websites are using the reputation system as structural mechanism for example:

a) Amazon's product review the most well-known example of object reputation for example the website asks:" Was this review helpful?" the reviewers program track those trusted review to provide context for potential buyer when it evaluate the potential of buyers.

b) EBay's feedback score is based on the number of transaction that completed by the buyer or seller and it's aggregated from thousand of individual transaction.

c) Xbox Live's achievement reward user when they reach the minor goals through the game and concurrently add this reward to community game score.
Table 1.2 shows that all of the top 25 websites on alexa.com that use at least one reputation system as critical part of their business.

Website	Vote to promote	Content rating and ranking	Content reviews and comments	Incentive karma (points)	Quality karma	Competitive karma	Abuse scoring
yahoo.*	††[a]	†††	†††	†	†	††	†††
google.*	††	†††	†	-	†	-	†††
youtube.com	†††[b]	††	†††	-	†	†	††
live.com	††	†††	††	†††	†	-	†††
facebook.com	†[c]	†††	††	††	†	†	†††
msn.com	†	†††	†††	-	†	††	††
wikipedia.org	†	-	††	-	†	-	†††
blogger.com	†	†	†††	-	†	-	†††
baidu.com	†	†††	†	†	†††	-	††
rapid-share.com	-[d]	-	-	†††	-	-	††
microsoft.com	-	-	†	-	-	-	†††
hi5.com	†	†	†	†	†	†	††

Table 2.2

Why we should use reputation system?

Reputation reporting systems have emerged important risk management mechanism in electronic communities. Reputation system collect, distribute, and aggregate feedback about the client past behavior. The goal of reputation system is to encourage trust-worthiness by using past behavior to predict the future behavior. So by aggregating this mechanism low quality transaction will replace by high quality and it's improving the whole quality of the system.

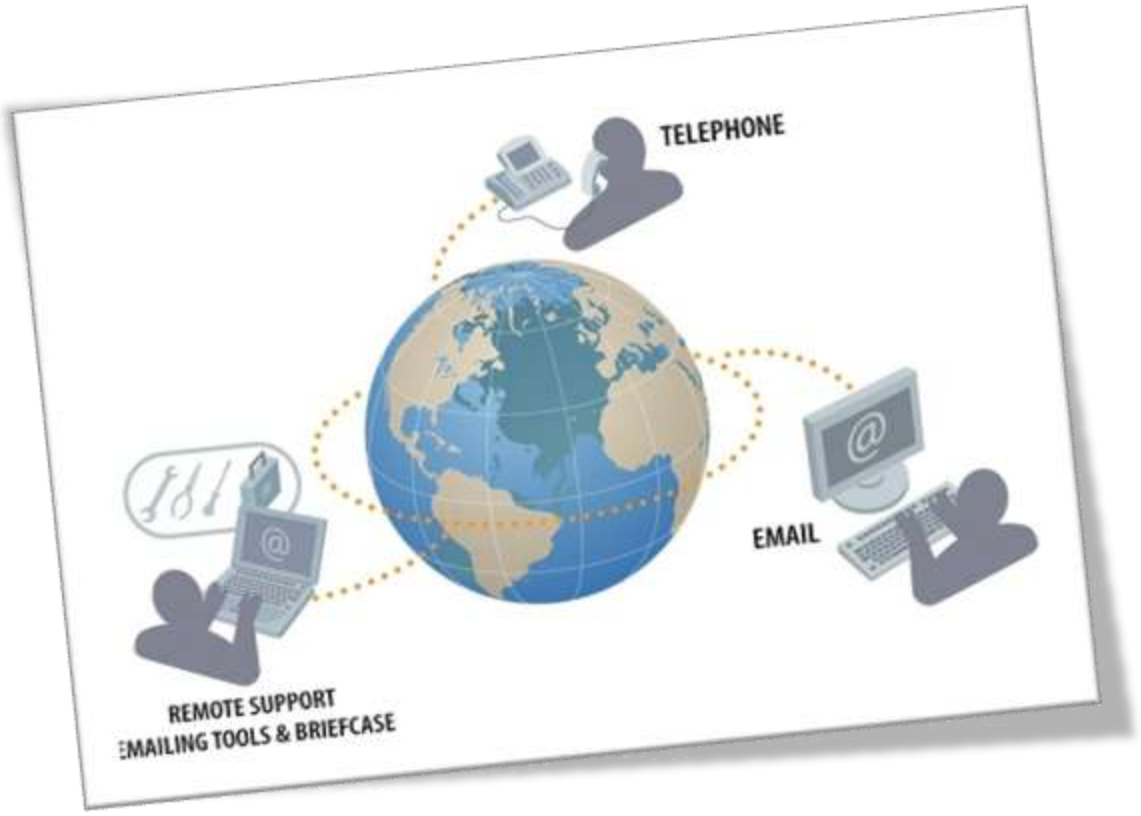
Why we need to design web reputation system?

Search engine are general purpose and implement ranking algorithm but beside of its effectiveness and efficiency most of the time do not meet user's expectation. User are normally dissatisfied with the result of choosing adjust information source of

search engine with given purpose. It's so obvious that to reach the best result, we need multiple information sources and not only the current algorithm that most of the search engine used these days.

We use possibility of adjusting the ranking provided by search engines with concern of web reputation of web information sources. The data quality define reputation as a dimension of information quality that measure the trustworthiness and important of information source. To define the data quality dimension we assess several metrics to show the impact of reputation of different search engine.

Till now we have discussed a lot of theory about setup and operate the reputation system. Now it's time to discuss the practical implementation of that.



3. Implementation

This part describes the project implementation for developing the reputation based selection of web information sources. The project implements PHP 5, JavaScript, and standard HTML.

The project will be capable of running on standard internet web browsers, although, the project is designed primarily around MS Internet Explorer. The interface for the project will provide a user the re-rank order of Google search engine with AHP methodology as a sort function and final result base on reputation system.

Project model:

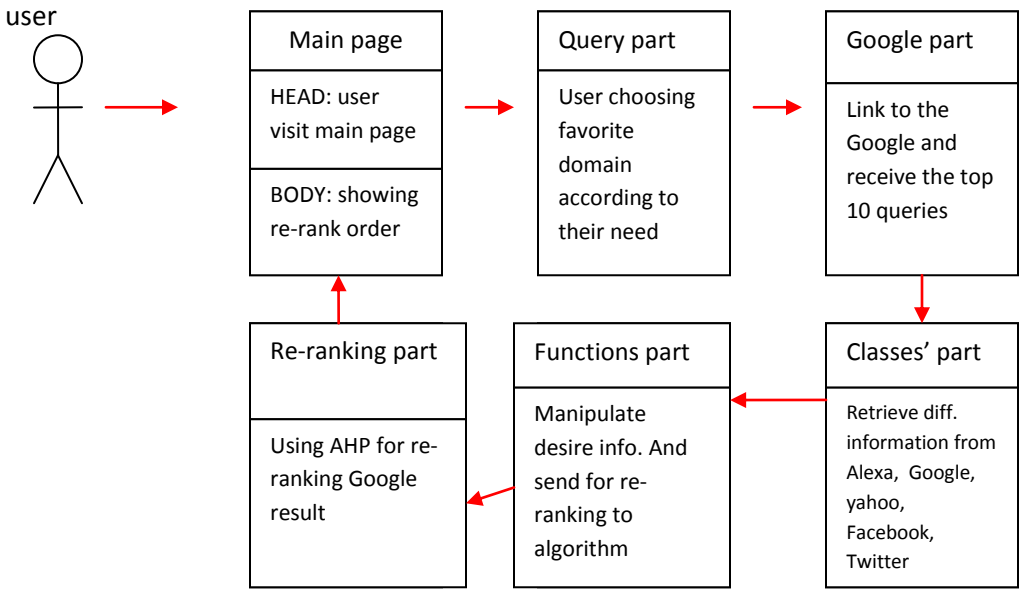

user
Main page
HEAD: user visit main page
BODY: showing re-rank order
Query part
User choosing favorite domain according to their need
Google part
Link to the Google and receive the top 10 queries
Classes' part
Retrieve diff. information from Alexa, Google, yahoo, Facebook, Twitter
Functions part
Manipulate desire info. And send for re-ranking to algorithm
Re-ranking part
Using AHP for re-ranking Google result

Figure 3

MAIN PAGE

Introduction:

The main page of the reputation based selection of web information sources is the entry point for all other pages contained in the website. The user will be able to begin querying, re-ranking, viewing different detail, viewing graphs or change user preferences from this point.

Details:

The main page will be developed in HTML / PHP/ JAVASCRIPT/SEOSTATES 2.01. The page will contain links to the other pages / modules. The layout of the page is based on page frames. Each frame will contain a link to a module. A query box for users that already enters to the site will be located in the middle side of the page.

JavaScript arrays:

The main page has 2 dimensional arrays that contain the values for query box by default for simplicity of user. The page will simply point to other pages that have other website API connections.

Error Handling:

The page will handle page errors by attempting to reload the module or report that the module is unavailable. If an invalid query is entered, the customer will be redirect to the main page and all the value will be rest.

- Reference: **index.php**

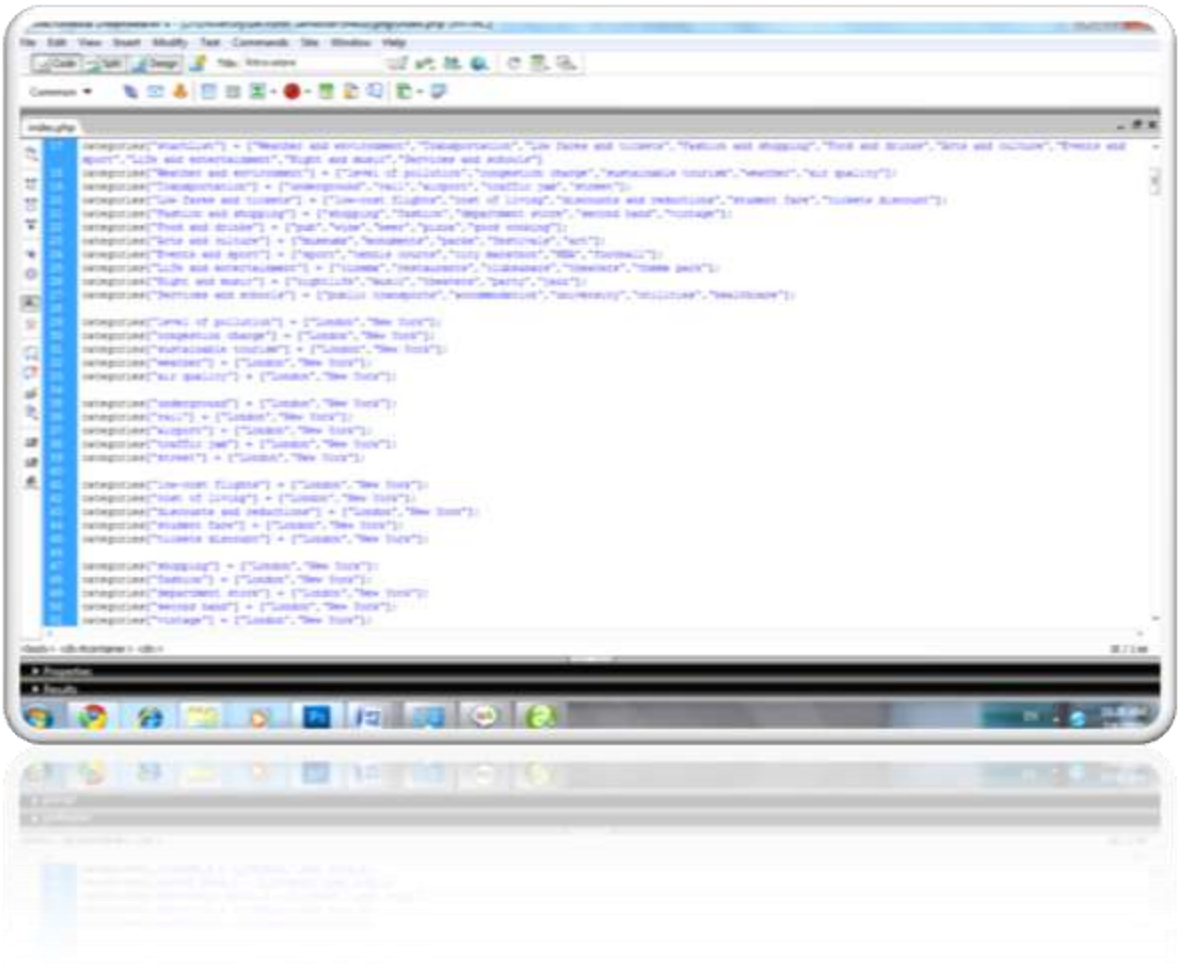


Figure 3.1

Query part

Introduction:

The query part will be the main part of for users to browse available domains and sub domain for viewing the result of chose query.

Details:

The page will query the database of Google's and extract the chose domain information. They chose domain information will then be store based on the domain and sub-domain in predefined array. This part will have a form box and 3 different combo boxes on the top middle of main page that will contain the main domain listing. The user can click on a domain to bring up the list of domains that are available.

If the domain contains sub-domain these will be displayed in the middle of the page and the user can proceed to drill down on the domains.

An option to show all domains will disregard the domains and sub- domains listings. When a user chooses to select an item for query the item ID will be passed to the query function module to retrieve the information and saved in array.

This page includes two main parts:

1) **Query form**
2) **Query function**

Error Handling:

In here, the errors could occur if the user selects too large of a quantity for queries. The software will have a limit based on execution time (120 sec). To resolve this issue define 8 numbers of query as default and don't let user choose number of query.

Information's stored in the array should be complete with details. When the information's are entered into the array by functions are not complete the error

handling will all be required. By making sure that all information is entered from the start the output will not contain partial information.

- Reference: **index.php**

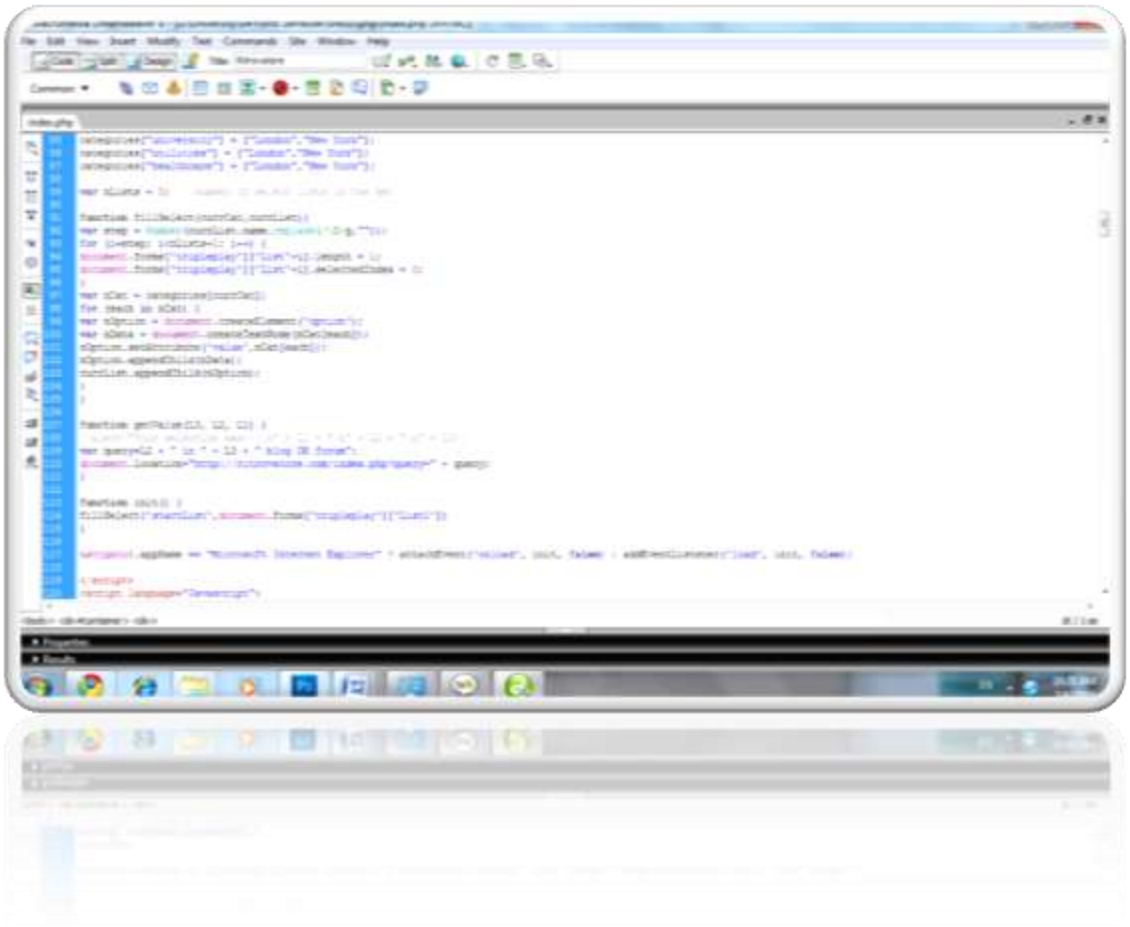


Figure 3.2

Query form and function:

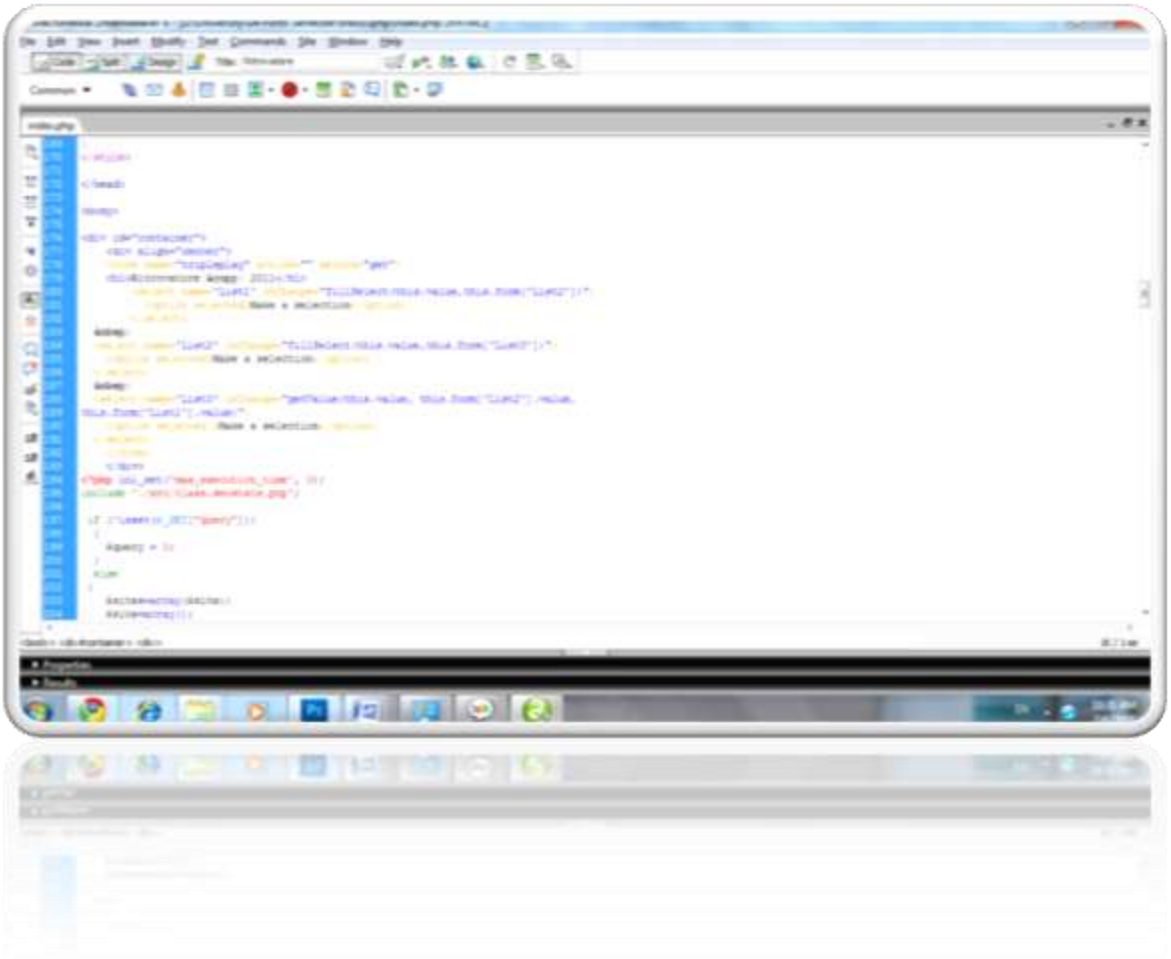


Figure 3.3

Retrieve part

Introduction:

The result of query part will be passing to this section to reorganize to store in JavaScript's arrays base on number of mentioned queries.

Details:

The query handling part will need to provide certain information to the retrieve part in order to store data in suitable way.

We design a for loop to read from handling part one by one and store in first dimension of array base on their 'name' and the detail of each query will be stored in second dimension of each array.

Error Handling:

Incomplete information will be dropped from the array. Only information that is complete will allow being store in array.

The page should check the values be submitted and determine if the value is null. Equality between the array key and the passed value will be check for preventing of function crash.

Retrieve function and array:

- Reference: **index.php**

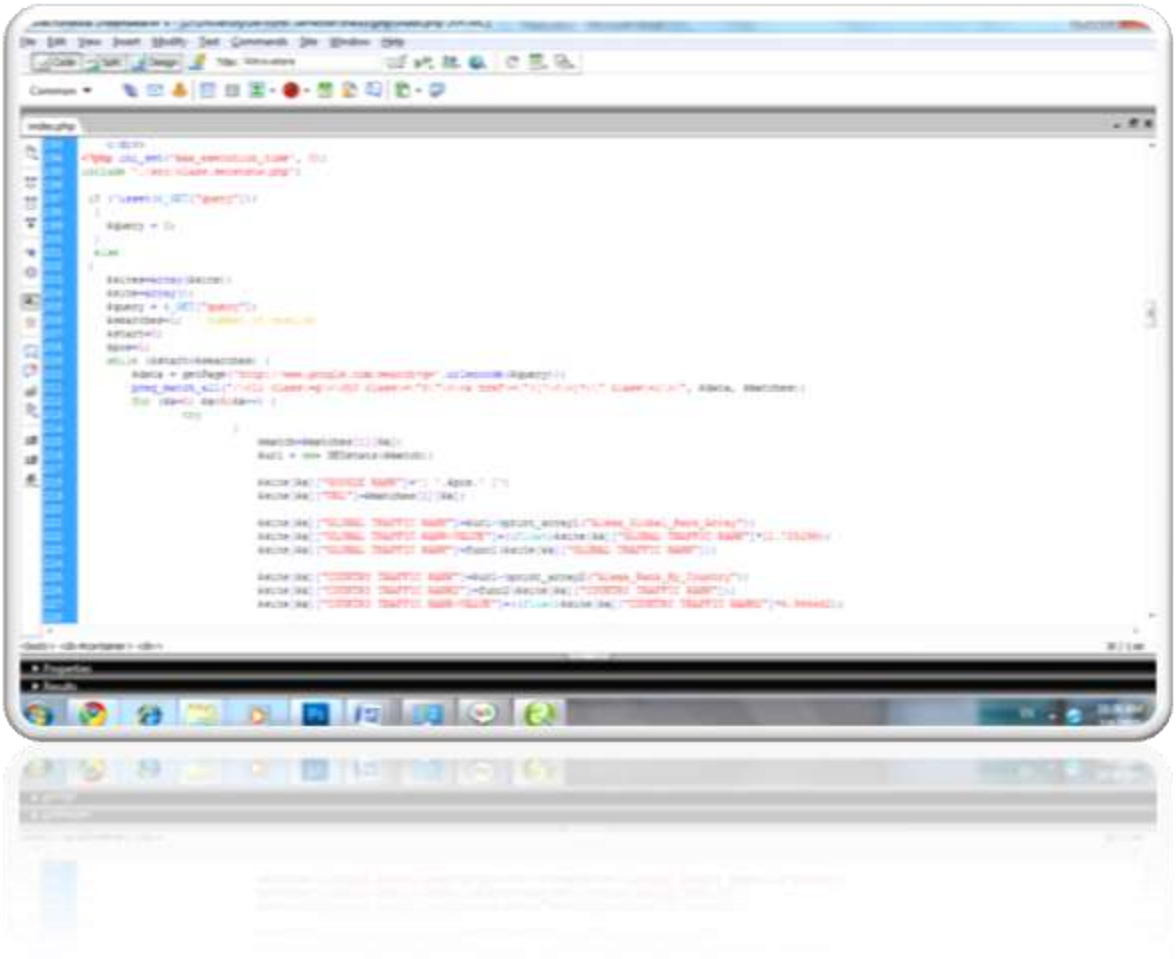


Figure 3.4

Class's part

Introduction:

The class's part is our most important part in the whole project. We gathered a lot of information from different data source likes: Alexa, Facebook, Twitter, Google and yahoo for each different query.

Details:

We should define a class to store different API user account info to make them allow retrieving information. For each different data source we make different classes that contain a lot of function to response and analysis our queries. We are listing below the name of each class and describe the function that each of them will do.

Classes:

- **Class.Config.php:**

  In this class we define different application ID of information source to retrieve information from that API.

- **Class.modules.php:**

  In this class we call different module form different class to enhance the speed of showing results of query.

- **Class.google.php:**

  In this class we implement different function to retrieve all related information from Google (for example: Google page speed score)

- **Class.yahoo.php:**

  In this class we implement different function to retrieve all related information from yahoo (for example: yahoo inbound link)

- **Class.alexa.php:**

  In this class we implement different function to retrieve all related information from Alexa (for example: Global ranking, country ranking)

- **Class.main.php:**

  In this class we gather all other classes functions result and manipulate final result that could be call in function part.

Error Handling:

To solving the all execution error we define a class that handle all the error from server side in any of data source

- Reference: **class.seostats.php**

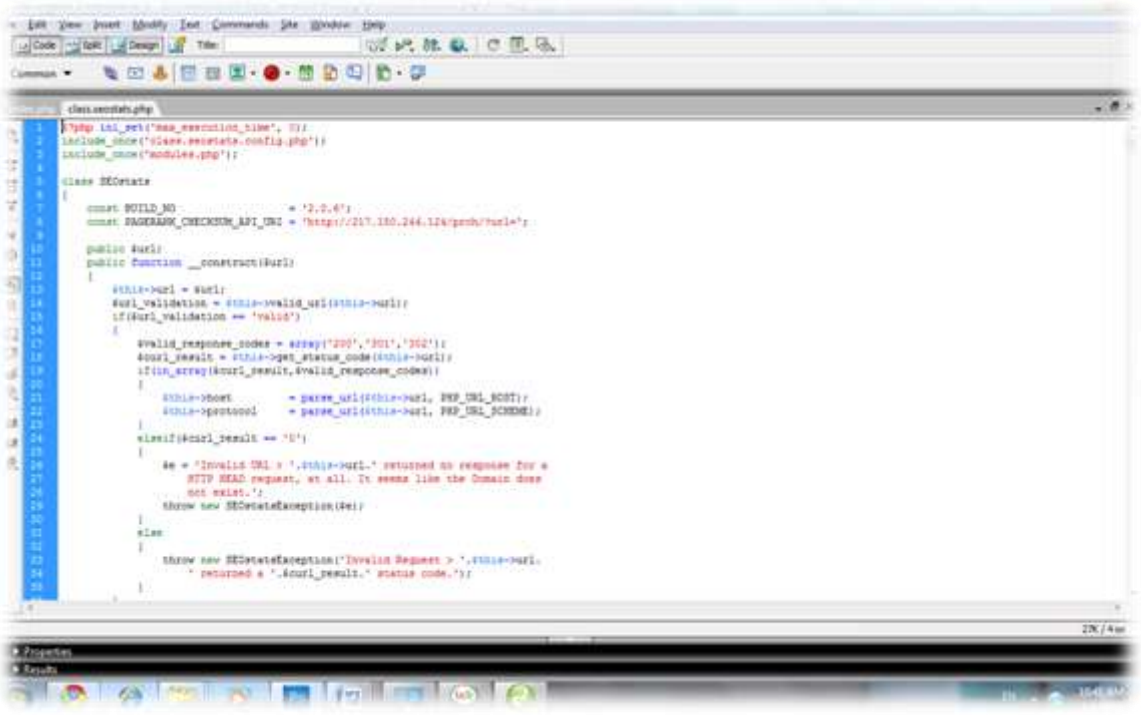


Figure 3.5

Function's part

Introduction:

The function's part is our project critical section because all of our calculation and print out visual effects are happening in this part. Also we standardized our metrics in this part.

Details:

Most of the retrieving results from classes are only the numbers and it's not user understandable. Also the visual of the result that will be appearing in the user screen should have some nice view that only functions can handle it.

One of the other main objective is when you load the different variables in JAVASCRIPT and you would like to pass them to the user screen the only way to show them is using function to make interconnection between these 2 parts.

To normalize the result of classes and make them standard we used different functions methodology to handle it.

Functions:

- **Screen popup function:**

  Alexa will provide for some of the metric graph ability that by using this function we open popup windows to show the result in graph ways.

- **Global rank function:**

  We classified global rank as below. As much as it has smaller we more weight to the site.

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
GLOBAL RANK	OTHER	X>50000 & X<100000	X>10000 & X<50000	X>500 & X<10000	X< 500

Table 3.1

- **Global country rank function:**

  We cannot standardize this metric because each country has different population and size. We show the website rank in most visited country.

- **Daily visit function:**

  First we find the website daily reach value, then after applying formula we will reach to the daily visit number. Then we classified the daily visitor as below.

  Daily visit number = ***(% daily reach*number of internet user)/100***

X	BAD	NORMAL	GOOD	PERFECT
DAILY VISIT	X>=0.0001 & X<0.001	X>=0.001 & X<0.01	X>=0.01 & X<0.1	X>=0.1 & X<1

Table 3.2

- **Time on site function:**

It all depends on the type of site and content you offer. You probably figured out on your own if you host 20 minute shows that people like watching like an anime show the average time on site would most likely be 20 minutes. But let's get out some hard facts from friends at SOURCE: CBS.

Based on 120M (120,000,000) Impressions, the Average time on the whole webpage is a whopping 33 Seconds!!!

So we classified the time on site base on the table shown below:

X	BAD	NORMAL	GOOD	PERFECT
TIME ON SITE (MIN)	X<0.55	X>=0.55 & X<1.55	X>=1.55 & X<3	X>=3

Table 3.3

- **Average load time function:**

  Information technology researches confirm that 75% of the Internet users do not return to sites that take longer than four seconds to load. They would suggest keeping your load time below 2 seconds.

  The Alexa website for this function has great information that contains all the user need. That's why we don't classify the result, and only we show the final return variable. It's comparing the average load time to the other internet website and release the % of how many websites have more or less average load time than this site.

- **Page speed score function:**

  Google Page Speed analysis using a Firefox browser measures how optimized the web page is in terms of loading time and provides a quantitative measurement that is known as a "Page Speed" score. This is a rating on a scale from 1 to 100. If a website scores 100, it means it is "perfectly optimized" for fast website loading.

  So we classified page speed score base on the table shown below:

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
PAGE SPEED SCORE	X<40	X>=40 & X<60	X>=60 & X<80	X>=80 & X<90	X>=90

Table 3.4

- **Facebook mentions function:**

  Facebook mention is the total amount of Facebook pages mention the URL. The last analysis shows that the web site mention is equal to 2,000 page views in 10 minutes. It shows that how much is important the number of mention as a metric in evaluating the website.

So we classified Facebook mention base on the table shown below:

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
FACEBOOK mention	X<1000	X>=1000 & X<10000	X>=10000 & X<100000	X>=100000 & X<1000000	X>=1000000

Table 3.5

- **Twitter mentions function:**

  Twitter mention is the total amount of Twitter pages mention the URL. The last analysis shows that the web site mention is equal to 2,000 page views in 10 minutes. It shows that how much is important the number of mention as a metric in evaluating the website.

  So we classified Facebook mention base on the table shown below:

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
Twitter mention	X<10	X>=10 & X<100	X>=100 & X<1000	X>=1000 & X<10000	X>=10000

Table 3.6

- **Alexa back-link function:**

Alexa back-link is a measure of Google's reputation. It means that the number of links to our specific website site from sites visited by users in the Alexa traffic panel. Those links that were not seen by users in the Alexa traffic panel are not counted. Multiple links from the same site are only counted once (ALEXA.com).

We classified the number of back-link base on the table that shown below:

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
ALEXA BACKLINK	X<100	X>=100 & X<1000	X>=1000 & X<10000	X>=10000 & X<100000	X>=100000

Table 3.7

- **Yahoo back-link function:**

Yahoo counts inbound links differently from every other search engine out there and we classified base on this.

X	BAD	NORMAL	GOOD	VERY GOOD	PERFECT
YAHOO BACKLINK	X<1000	X>=1000 & X<10000	X>=10000 & X<100000	X>=100000 & X<1000000	X>=1000000

Table 3.8

- **Bounce rate function:**

If you experience high bounce rates (over 70%) then it might be that your website stinks, to put it bluntly, and you need to redesign it or do a better job communicating through your entry pages.

To sum up, the bounce rate is affected by many things and there is no blanket answer that can be applied to all website to say "When your bounce rate is high you should…" Each website is different, each situation is different, and analyzing bounce rate data requires a hands-on approach–just one more reason why SEO can never be fully automated (theorganicseo.com).

So we classified the bounce rate base on the following table:

X	BAD	NORMAL	GOOD	PERFECT
BOUNCE RATE	X>=70	X>=50 & X<70	X>=40 & X<50	X<40

Table 3.9

- **Page view per user function:**

The number of page views per user is a key indicator of the quality, depth and breadth of content on a given website (Alexa.com).Average page views per visit are an excellent indicator of how compelling and easily navigated your content is.

We calculate it by the total number of page views divided by the total number of visits during the same timeframe.

Page Views / Visits = Average Page Views per Visit

Then we classified the result base on the table shown below:

X	BAD	NORMAL	GOOD	PERFECT
PAGE VIEW PER USER	X>=2 & X<3	X>=2 & X<3	X>=3 & X<5	X>=5

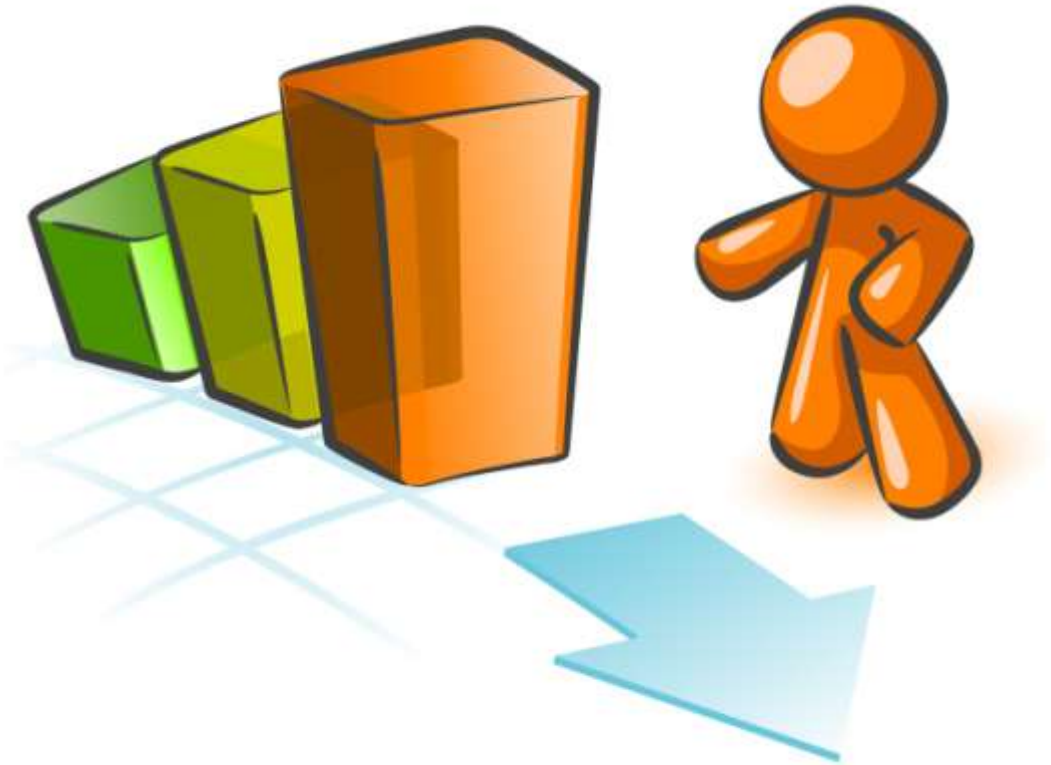
Table 3.10

- **Search visit function:**

This function calculate the number of visitor from different search engine find us and move over our website. There isn't any good range for classification of these data but according to the research's we figure out that we can classify results by the number of visit per day. So we use this indicator as a key point for defining the following table.

X	BAD	NORMAL	GOOD	PERFECT
SEARCH VISIT	X>=0 & X<18	X>=18 & X<30	X>=30 & X<50	X>=50

Table 3.11



Re-ranking part

Introduction:

In this part of our project our main goal was re-ranking the Google result base on our metrics. For re-ranking there are several methods called MCDA (Multi criteria decision analysis)

There are many MCDA / MCDM methods in use today. However, often different methods may yield different results for exactly the same problem. In other words, when exactly the same problem data are used with different MCDA / MCDM methods, such methods may recommend different solutions even for very simple problems (i.e., ones with very few alternatives and criteria). (wikipedia.org)

The choice of which model is most appropriate depends on the problem at hand and may be to some extent dependent on which model the decision maker is most comfortable with. We choose AHP (Analytic hierarchy process) to make decision for our final re-ranking.

What is the AHP and how did we apply AHP to our model?

The Analytic Hierarchy Process (AHP) is a structured technique for dealing with complex decisions.

Rather than prescribing a "correct" decision, the AHP helps decision makers find one that best suits their goal and their understanding of the problem—it is a process of organizing decisions that people are already dealing with, but trying to do in their heads.(wikipedia.org).

Our AHP steps of applying:

1) *We model our metrics in a structured tree hierarchy and define the alternatives for reaching our model's goal. We had four main domain criteria and we locate our predefined metrics under the related domain and reach to 13 sub-domain criteria.*

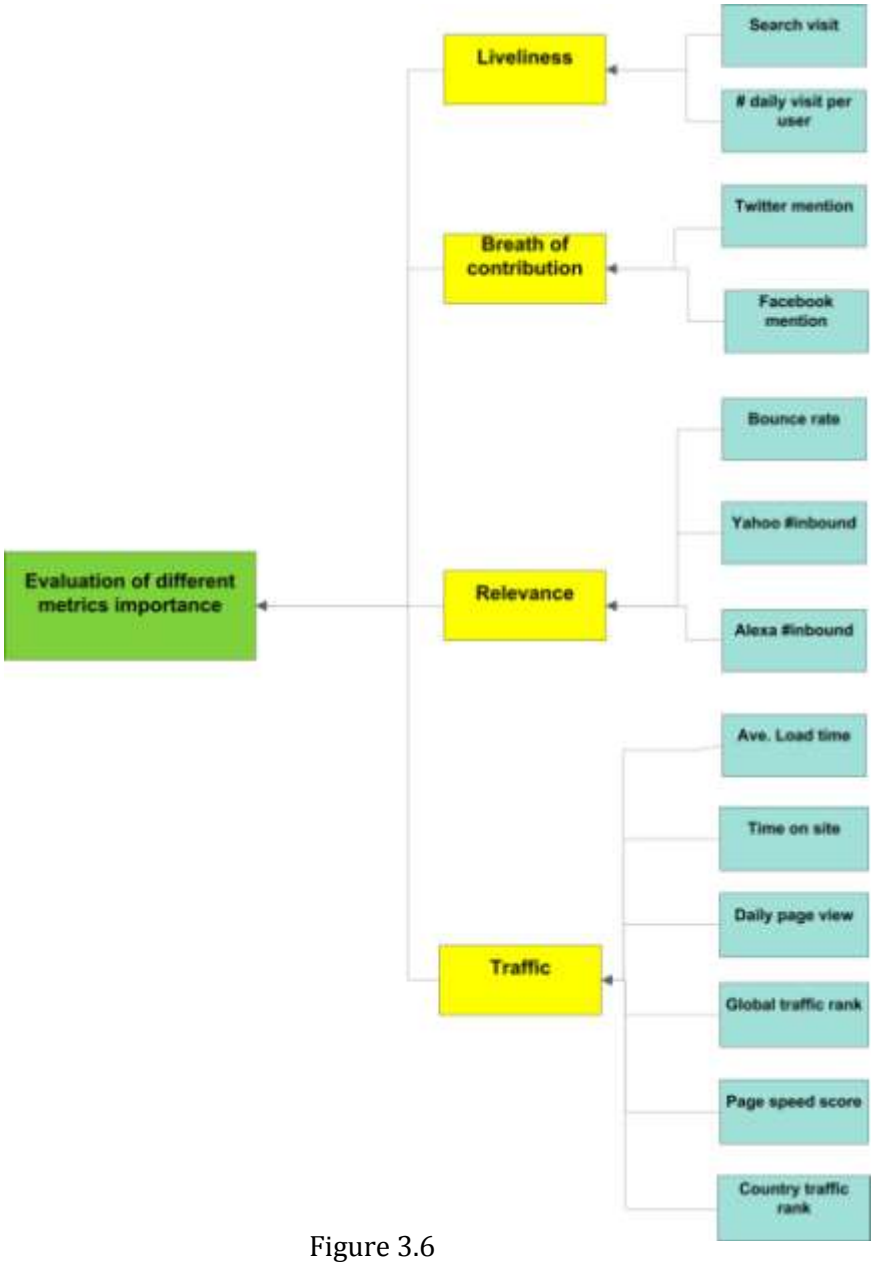


Figure 3.6

2) *We comparing the importance of each metric to give the priorities among the parent element of our hierarchy based on our goal. We gave our priorities mostly base on the number of sub-domain.*

For example as we had six sub-domains in traffic part we gave higher priorities to traffic rather than others.

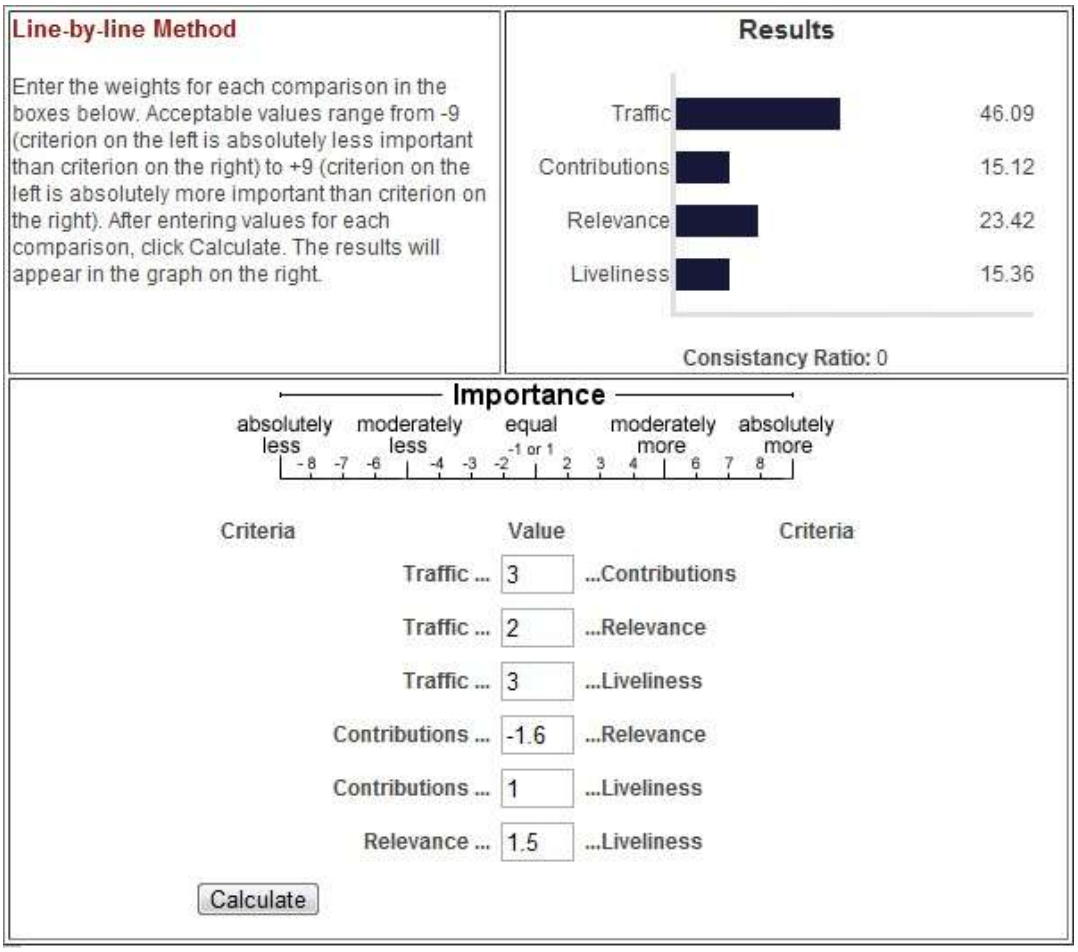

Line-by-line Method
Enter the weights for each comparison in the boxes below. Acceptable values range from -9 (criterion on the left is absolutely less important than criterion on the right) to +9 (criterion on the left is absolutely more important than criterion on the right). After entering values for each comparison, click Calculate. The results will appear in the graph on the right.
Results
Traffic 46.09
Contributions 15.12
Relevance 23.42
Liveliness 15.36
Consistancy Ratio: 0
Importance
absolutely less
moderately less
equal
-1 or 1
moderately more
absolutely more
-8 -7 -6 -4 -3 -2 2 3 4 6 7 8
Criteria
Value
Criteria
Traffic ... 3 ...Contributions
Traffic ... 2 ...Relevance
Traffic ... 3 ...Liveliness
Contributions ... -1.6 ...Relevance
Contributions ... 1 ...Liveliness
Relevance ... 1.5 ...Liveliness
Calculate

Figure 3.7

3) *The same as step above. We gave priorities to each sub-domain according to its importance role in our final decision.*

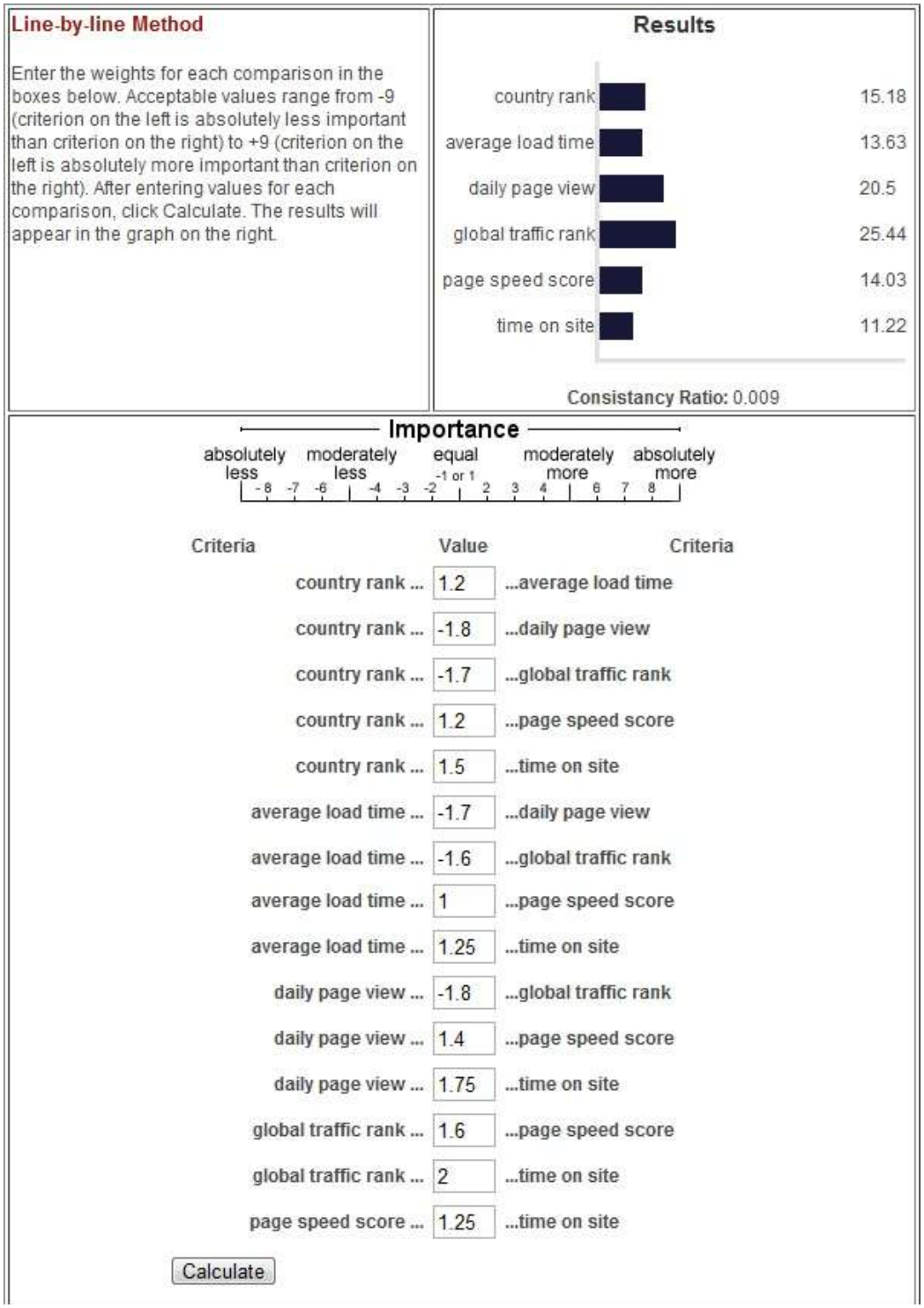

Line-by-line Method
Enter the weights for each comparison in the boxes below. Acceptable values range from -9 (criterion on the left is absolutely less important than criterion on the right) to +9 (criterion on the left is absolutely more important than criterion on the right). After entering values for each comparison, click Calculate. The results will appear in the graph on the right.
Results
country rank 15.18
average load time 13.63
daily page view 20.5
global traffic rank 25.44
page speed score 14.03
time on site 11.22
Consistancy Ratio: 0.009
Importance
absolutely less
moderately less
equal
-1 or 1
moderately more
absolutely more
- 8 -7 -6 -4 -3 -2 2 3 4 6 7 8
Criteria
Value
Criteria
country rank ... 1.2 ...average load time
country rank ... -1.8 ...daily page view
country rank ... -1.7 ...global traffic rank
country rank ... 1.2 ...page speed score
country rank ... 1.5 ...time on site
average load time ... -1.7 ...daily page view
average load time ... -1.6 ...global traffic rank
average load time ... 1 ...page speed score
average load time ... 1.25 ...time on site
daily page view ... -1.8 ...global traffic rank
daily page view ... 1.4 ...page speed score
daily page view ... 1.75 ...time on site
global traffic rank ... 1.6 ...page speed score
global traffic rank ... 2 ...time on site
page speed score ... 1.25 ...time on site
Calculate

Figure 3.8

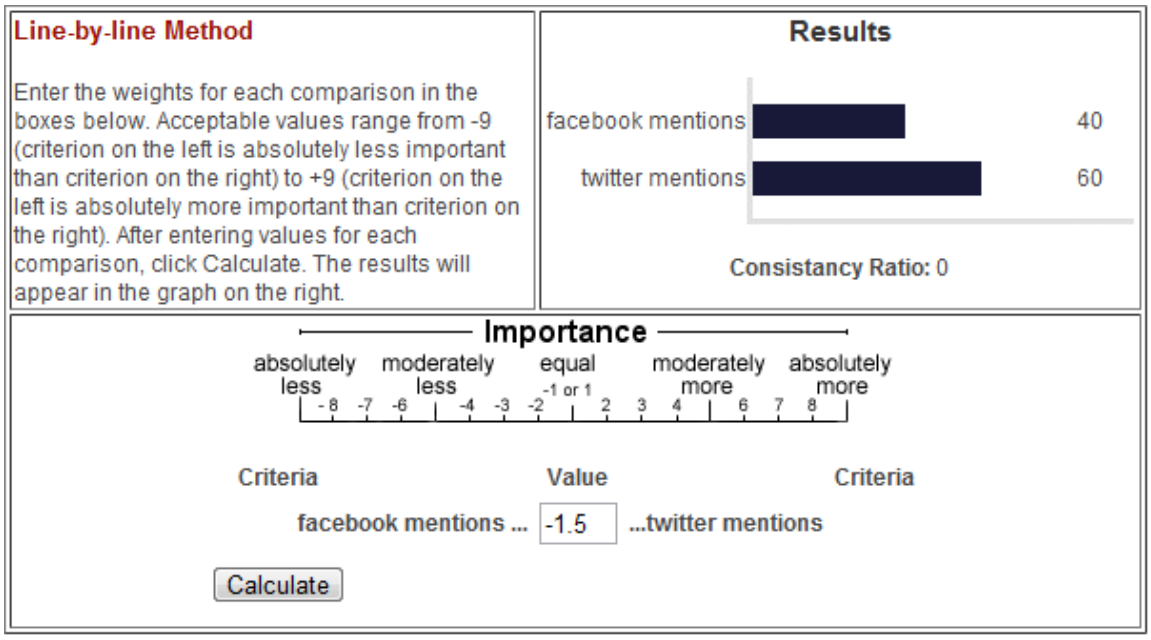

Line-by-line Method
Enter the weights for each comparison in the boxes below. Acceptable values range from -9 (criterion on the left is absolutely less important than criterion on the right) to +9 (criterion on the left is absolutely more important than criterion on the right). After entering values for each comparison, click Calculate. The results will appear in the graph on the right.
Results
facebook mentions
40
twitter mentions
60
Consistancy Ratio: 0
Importance
absolutely less
moderately less
equal
-1 or 1
moderately more
absolutely more
- 8 -7 -6 -4 -3 -2 2 3 4 6 7 8
Criteria
Value
Criteria
facebook mentions ...
-1.5
...twitter mentions
Calculate

Figure 3.4

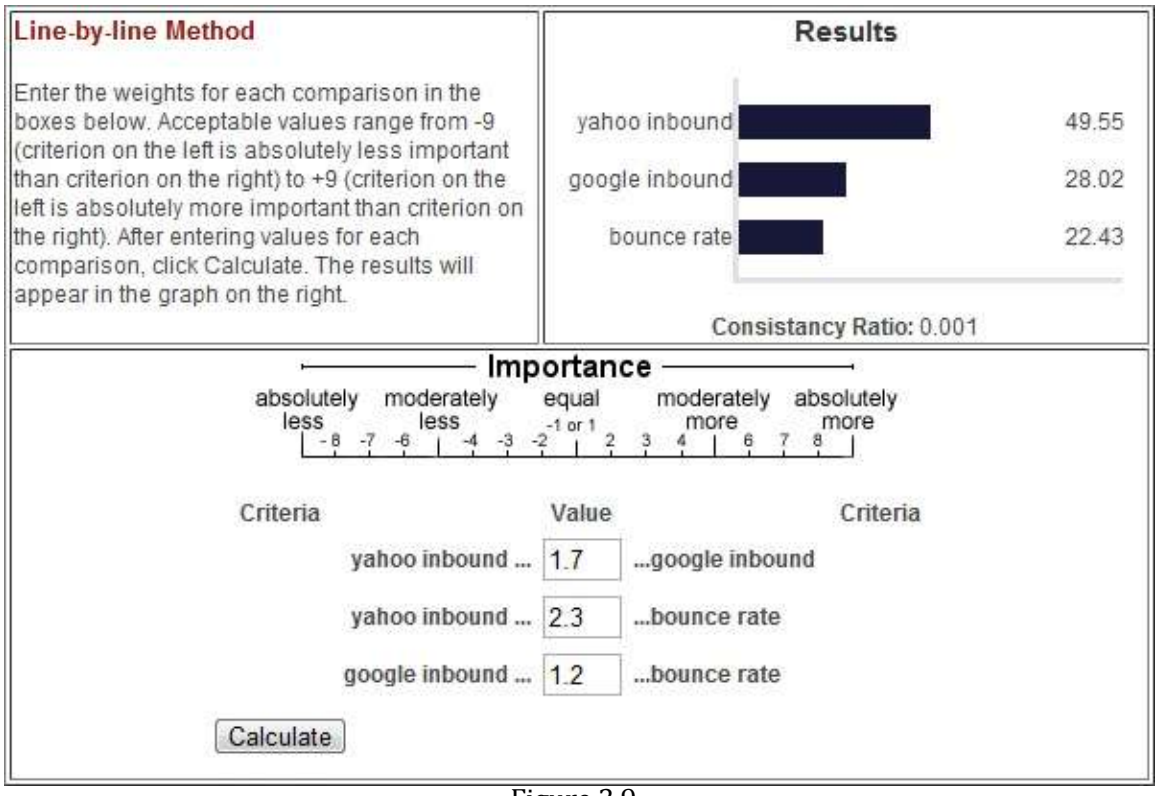

Line-by-line Method
Enter the weights for each comparison in the boxes below. Acceptable values range from -9 (criterion on the left is absolutely less important than criterion on the right) to +9 (criterion on the left is absolutely more important than criterion on the right). After entering values for each comparison, click Calculate. The results will appear in the graph on the right.
Results
yahoo inbound
49.55
google inbound
28.02
bounce rate
22.43
Consistancy Ratio: 0.001
Importance
absolutely less
moderately less
equal
-1 or 1
moderately more
absolutely more
- 8 -7 -6 -4 -3 -2 2 3 4 6 7 8
Criteria
Value
Criteria
yahoo inbound ...
1.7
...google inbound
yahoo inbound ...
2.3
...bounce rate
google inbound ...
1.2
...bounce rate
Calculate

Figure 3.9

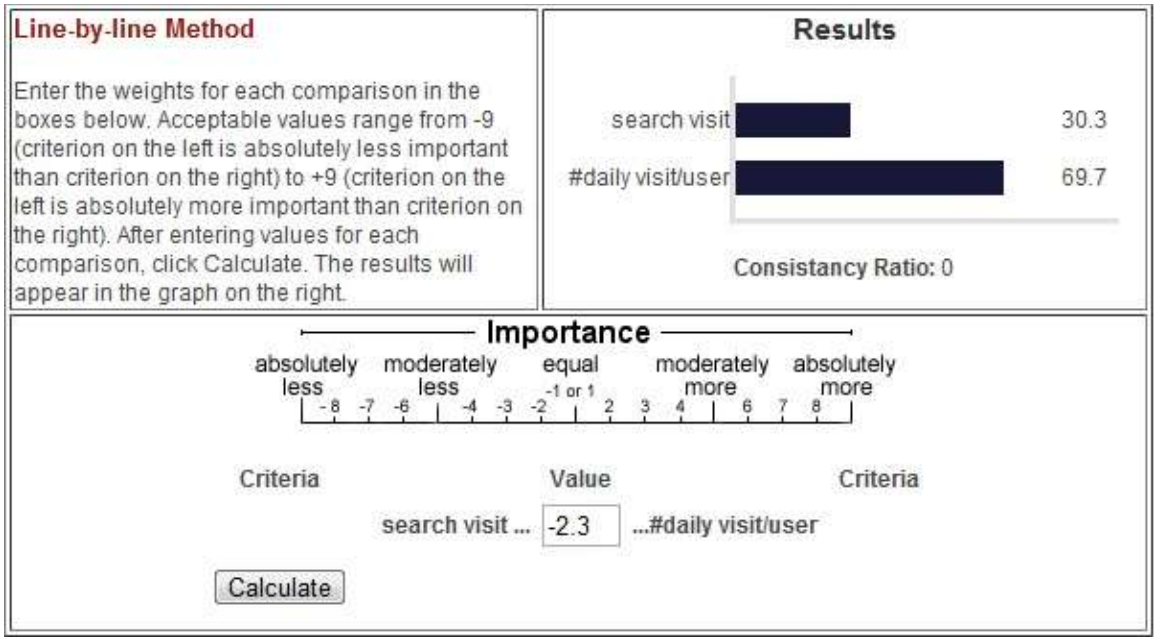

Line-by-line Method
Enter the weights for each comparison in the boxes below. Acceptable values range from -9 (criterion on the left is absolutely less important than criterion on the right) to +9 (criterion on the right is absolutely more important than criterion on the left). After entering values for each comparison, click Calculate. The results will appear in the graph on the right.
Results
search visit
30.3
#daily visit/user
69.7
Consistancy Ratio: 0
Importance
absolutely less
moderately less
equal
-1 or 1
moderately more
absolutely more
-8 -7 -6 -4 -3 -2 2 3 4 6 7 8
Criteria
Value
Criteria
search visit ...
-2.3
...#daily visit/user
Calculate

Figure 3.10

4) Then we check the consistency of the judgments in our project

Figure: the weight of each sub-domain

Figure 3.11

Figure: calculating the weight of each sub-domain

Traffic			
country rank	15.18	46.09	6.996462
average load time	13.63		6.282067
daily page view	20.5		9.44845
global traffic rank	25.44		11.725296
page speed score	14.03		6.466427
time on site	11.22		5.171298
Contribution			
mention facebook	40	15.13	6.052
mention twitter	60		9.078
Relevance			
#inbound link yahoo	49.55	23.42	11.60461
#inbound link google	28.02		6.562284
bounce rate	22.43		5.253106
liveliness			
search visit	30	15.36	4.608
#daily view per user	70		10.752

Figure 3.12

5) Make the final decision

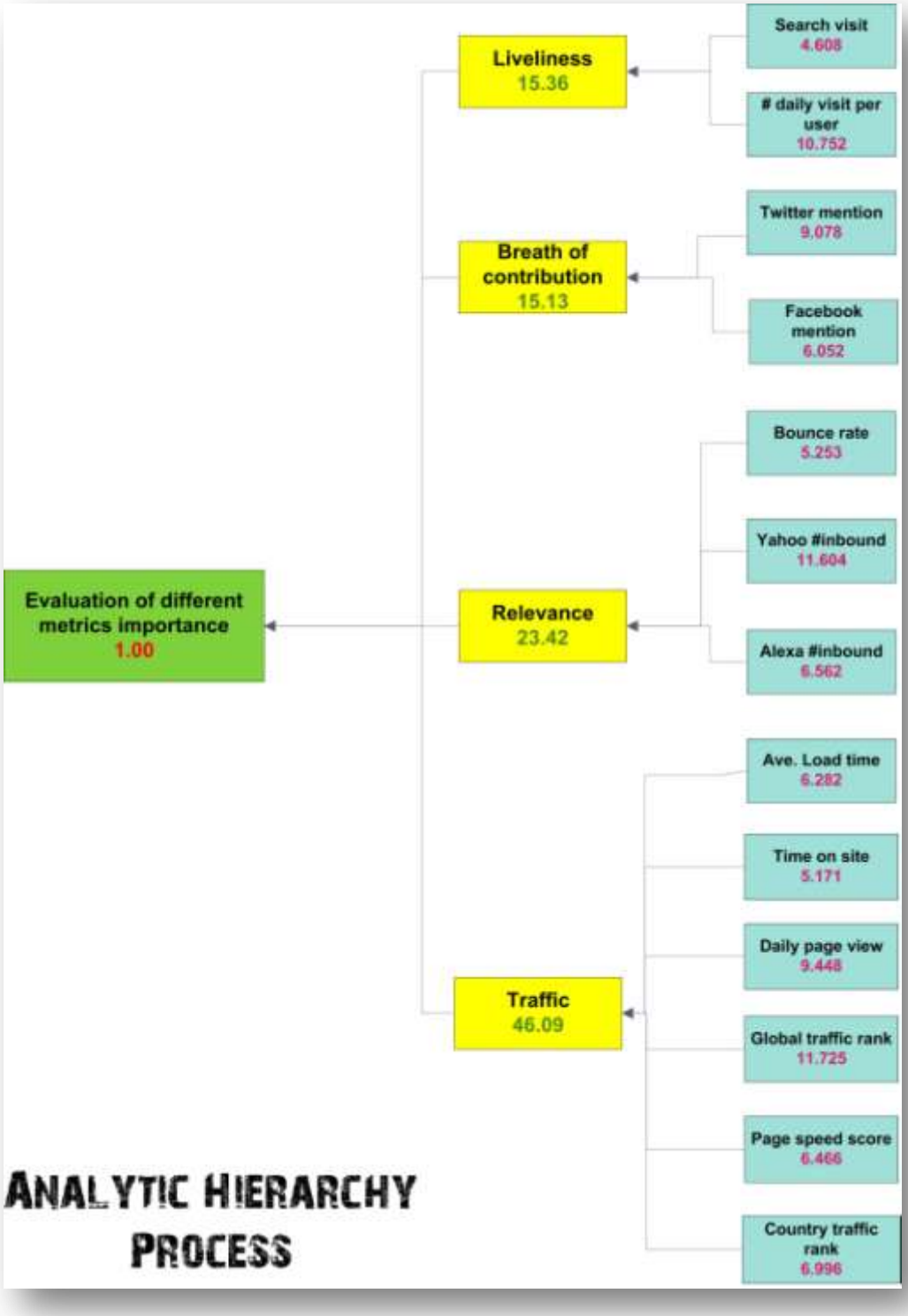


Figure 3.13

AHP variable:

1) ***Global-traffic rank:***

We consider this variable as a negative factor because when the global rank is higher it shows the site value is less.

```
$site[$x]["GLOBAL TRAFFIC RANK-VALUE"]=((float)$site[$x]["GLOBAL TRAFFIC
RANK"]*11.725296);
```

2) ***Country traffic-rank:***

We consider this variable as a negative factor because when the global rank is higher it shows the site value is less.

```
$site[$x]["COUNTRY TRAFFIC RANK-VALUE"]=((float)$site[$x]["COUNTRY
TRAFFIC RANK2"]*6.996462);
```

3) ***Daily visit:***

We consider this variable as a positive factor because when the daily visit is higher it shows the site value is more interesting for the users.

```
$site[$x]["DAILY VISIT-VALUE"]=((float)$site[$x]["DAILY
VISIT2"]*9.44845);
```

4) ***Time on site:***

We consider this variable as a positive factor because when the time on site is higher it shows the user is more comfortable and interest to spend time on that site.

```
$site[$x]["TIME ON SITE-VALUE"]=((float)$site[$x]["TIME ON
SITE"]*5.171298);
```

5) ***Average load time:***

We consider this variable as a negative factor because when the average load time is higher it shows that the site need much more time to be loaded and it decrease the interest of the user to visit this site next time.

```
$site[$x]["AVERAGE LOAD TIME-VALUE"]=((float)$site[$x]["AVERAGE LOAD
TIME2"]*6.282067);
```

6) ***Page speed score:***

We consider this variable as a positive factor (this number is measured by Google and its range is between 0-100)

```
$site[$x]["PAGE SPEED SCORE-VALUE"]=((float)$site[$x]["PAGE SPEED
SCORE"]*6.466427);
```

7) ***Facebook mentions:***

we consider this variable as a positive factor because when the site mentioned in Facebook or any social network has a direct effect on the increasing of the visitor in that website.

```
$site[$x]["FACEBOOK MENTION-VALUE"]=((float)$site[$x]["FACEBOOK
MENTION"]*6.052);
```

8) ***Twitter mentions:***

we consider this variable as a positive factor because when the site mentioned in Twitter or any social network has a direct effect on the increasing of the visitor in that website.

```
$site[$x]["TWITTER MENTION-VALUE"]=((float)$site[$x]["TWITTER
MENTION"]*9.078);
```

9) ***Alexa inbound links:***

We consider this variable as a positive factor because when the inbound links is higher it shows the site has more connection with other sites. This is one of the newest factors in web reputation techniques that such a website like Alexa and yahoo try to provide this information for users.

```
$site[$x]["ALEXA BACKLINK-VALUE"]=((float)$site[$x]["ALEXA
BACKLINK"]*6.562284);
```

10) ***Yahoo inbounds links:***

We consider this variable as a positive factor because when the inbound links is higher it shows the site has more connection with other sites. This is one of the newest factors in web reputation techniques that such a website like Alexa and yahoo try to provide this information for users

```
$site[$x]["YAHOO BACKLINK-VALUE"]=((float)$site[$x]["YAHOO
BACKLINK"]*11.60461);
```

11) Bounce rate:

We consider this variable as a negative factor because when the bounce rate is higher it means that no-one goes to that website regularly. When the user enters to the site he/she just view the first page and not interested in going deeper in that site.

```
$site[$x]["BOUNCE RATE-VALUE"]=((float)$site[$x]["BOUNCE
RATE"]*5.253106);
```

12) Page view per user:

We consider this variable as a positive factor because when the page view per user is high its shows this site is more interesting and valuable for users.

```
$site[$x]["PAGE VIEW PER USER-VALUE"]=((float)$site[$x]["PAGE VIEW PER
USER"]*10.752);
```

13) Search visit:

We consider this variable as a positive factor because when the search visit is high its shows mostly user find us in search engines.

```
$site[$x]["SEARCH VISIT-VALUE"]=((float)$site[$x]["SEARCH VISIT"]*4.608);
```

14) AHP value:

We sum up the entire above variable according to their factor and we re-rank our top 10 based on this value.

```
$site[$x]["AHP VALUE"]=round($site[$x]["DAILY VISIT-VALUE"]-
$site[$x]["GLOBAL TRAFFIC RANK-VALUE"]-$site[$x]["COUNTRY TRAFFIC RANK-
VALUE"]+$site[$x]["TIME ON SITE-VALUE"]-$site[$x]["AVERAGE LOAD TIME-
VALUE"]+$site[$x]["PAGE SPEED SCORE-VALUE"]+$site[$x]["FACEBOOK MENTION-
VALUE"]+$site[$x]["TWITTER MENTION-VALUE"]+$site[$x]["ALEXA BACKLINK-
VALUE"]+$site[$x]["YAHOO BACKLINK-VALUE"]-$site[$x]["BOUNCE RATE-
VALUE"]+$site[$x]["PAGE VIEW PER USER-VALUE"]+$site[$x]["SEARCH VISIT-
VALUE"],3);
```

AHP functions:

We re-ranking the Google top 10 result bases on AHP-value that calculated above. As the AHP value is higher we show the website in the top of re-ranked list.

4. CONCLUSION

This thesis has presented the reputation-based ranking and the results of the analyses that were conducted to identify the relevance of data quality and reputation metrics over the Google ranking algorithm and to measure the distance between the Google's ranking and the reputation based ranking.

Results show that the reputation metrics have different relevance to Google ranking algorithm since each ranking that is based along each of the reputation metrics has different distance values when comparing it with the Google's ranking.

Moreover, there is a difference (distance) between the Google's ranking and the ranking that is based on the reputation metrics (reputation based ranking), and the difference percentage values show that the distance between the Google's ranking and the reputation based ranking is significant and becomes less when taking into consideration the reputation metrics weights.

The reputation metrics can be ordered according to their relevance to the Google ranking algorithm as follow starting from the one that has the most relevance: On line since, Daily pageviews, Traffic Rank, In bound links, Bounce rate, Average number of comments to post within 24 hours, Number of open discussion/post per day, Time on site, Daily pageviews/user, Number of distinct tags and finally Number of comments to selected post;

The reputation based ranking of the information sources and the assessment of the quality of their information can improve the selection of information sources and can help Web users to select the most authoritative sources since it takes in to account the effective interaction between the users and the information sources in its ranking algorithm.

This is especially relevant in the context of the market monitoring, where Web users not only retrieve and access Web resources to get an idea about a key interest topic, but also to take some kind of choice/decision.

Our online reputation system should not be limited to what we do not want folk to see or say about us. Let the understanding that you're in public guide your judgment about what to post on the web. After independent analysis, we are persuaded that it offers better protection and performance for search engines if they using web reputation for selecting of information sources.

5. USER MANUAL

Using the web reputation based selection of information source website is simple. The site is free and provides:

- *A summary of available metrics in web reputation system*
- *A checklist in each domain area which you can use to have different query based on your need*
- *A view of our new re-rank Google order to better understanding of web-reputation based selection of different information source*

❖ *Getting started:*

- *Anyone can view the web reputation based selection website by accessing* ***www.ritrovatore.com*** *but to complete drop down list and contribute to our progress first you should read the report to understand each domain and sub-domains to get a related query to see your desire result.*

Figure 5.1

You can read all the web-reputation material on the site without having to log in, but to use the result of query you should wait around 1 minute.

❖ *Browsing information:*

- *The web-reputation website information on the site in organized into 3 main drop down list menu that each of them is dependant to the others:*

1) Domains:

The first drop down list that browsing10 main domain which we are explained before and all of our queries are based on this 10 main domain.

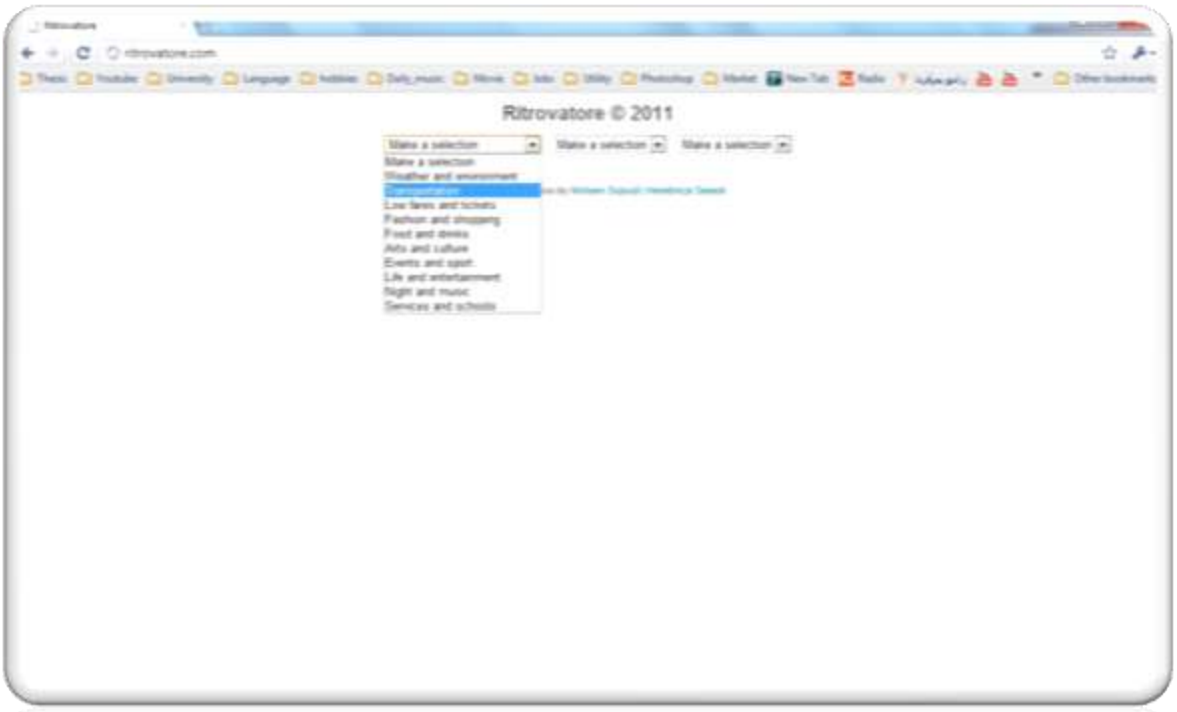


Figure 5.2

2) Sub-domains:

After selecting one of these domains the next combo box will change automatically and brings the sub-domains related to the first domain that already chose.

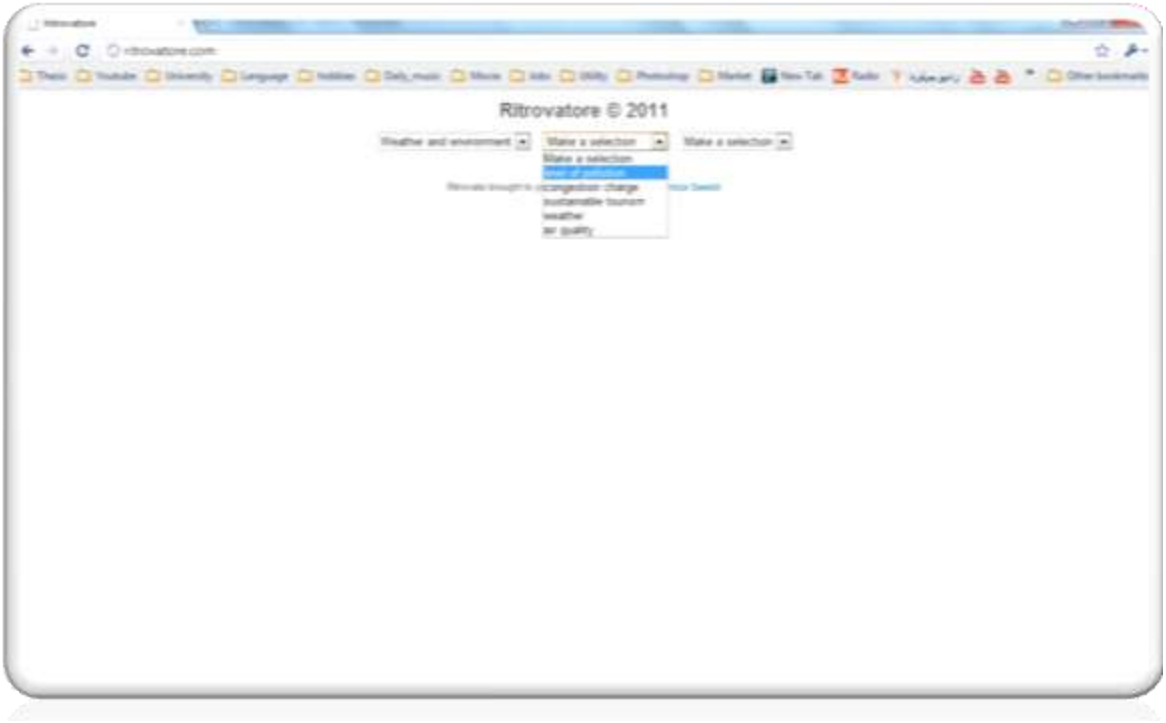


Figure 5.3

3) Cities:

After selecting the sub-domains the third combo-box shows the related city based on 2 previous selected options. By clicking on the last combo-box we should wait around 1 min to servers bring up the whole information about our queries

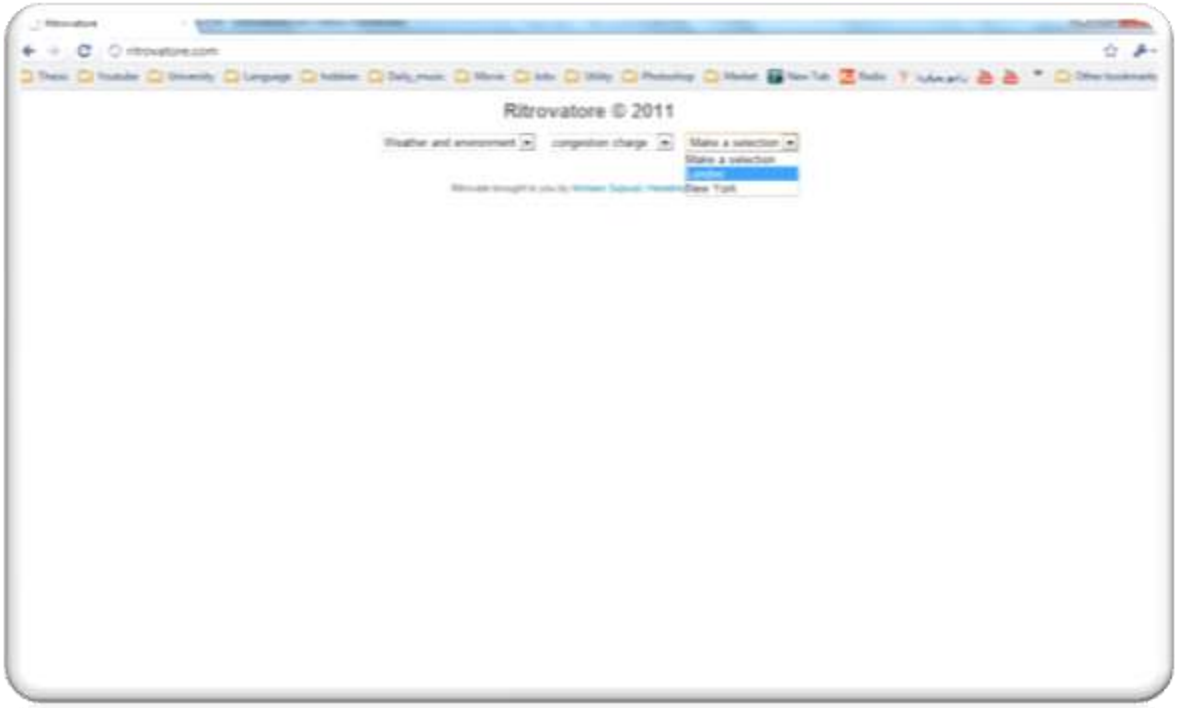


Figure 5.4

❖ *Query's result*

The next step is to understanding what the query result that will be shown on the screen of the user is. We classified the result into 2 main parts:

1) Google rank:

- *The top list is the first eight result of the Google based on the selected queries.*

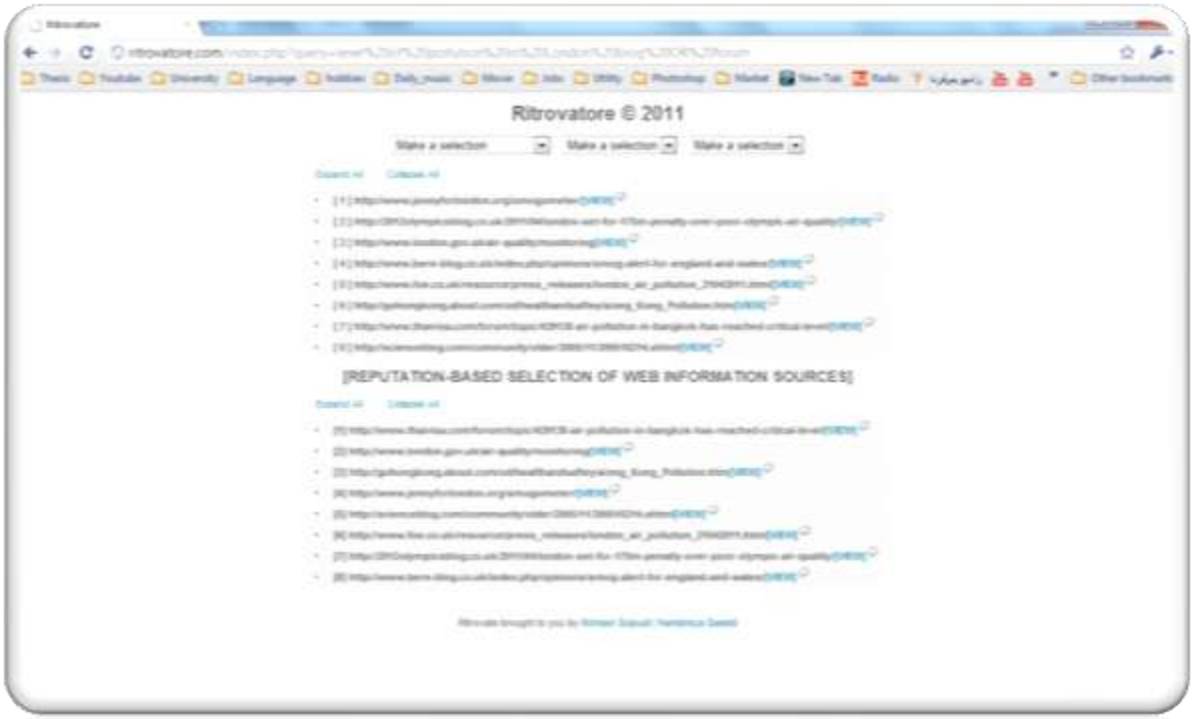


Figure 5.5

At the end of each line there is a blue button called "VIEW" that by clicking over on it we will browsed to that website and also if you move over this button the snap shot of the website will appears.

Figure 5.6

At the beginning of each line there is a "+" and "-" button that by clicking over these button you can expand or collapse the information related to that specified URL.

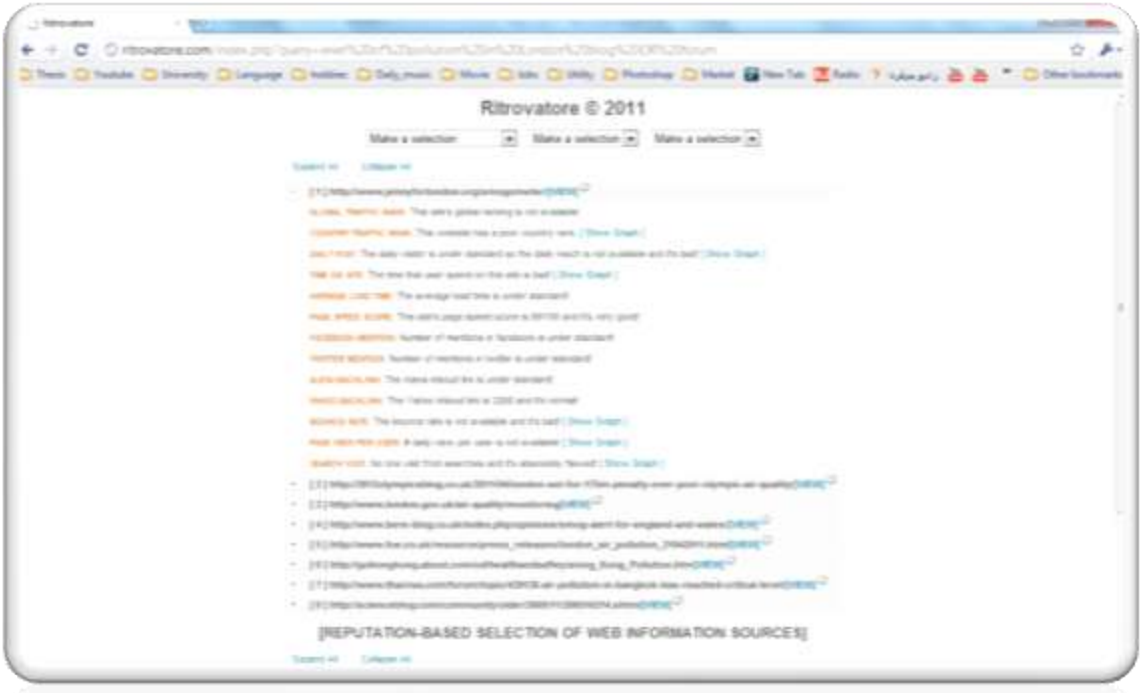


Figure 5.7

For easier surfing of the page we placed 2 button on the of each list name as " expand all" and "collapse all" that by clicking over those button the whole list will be expanded/collapsed.

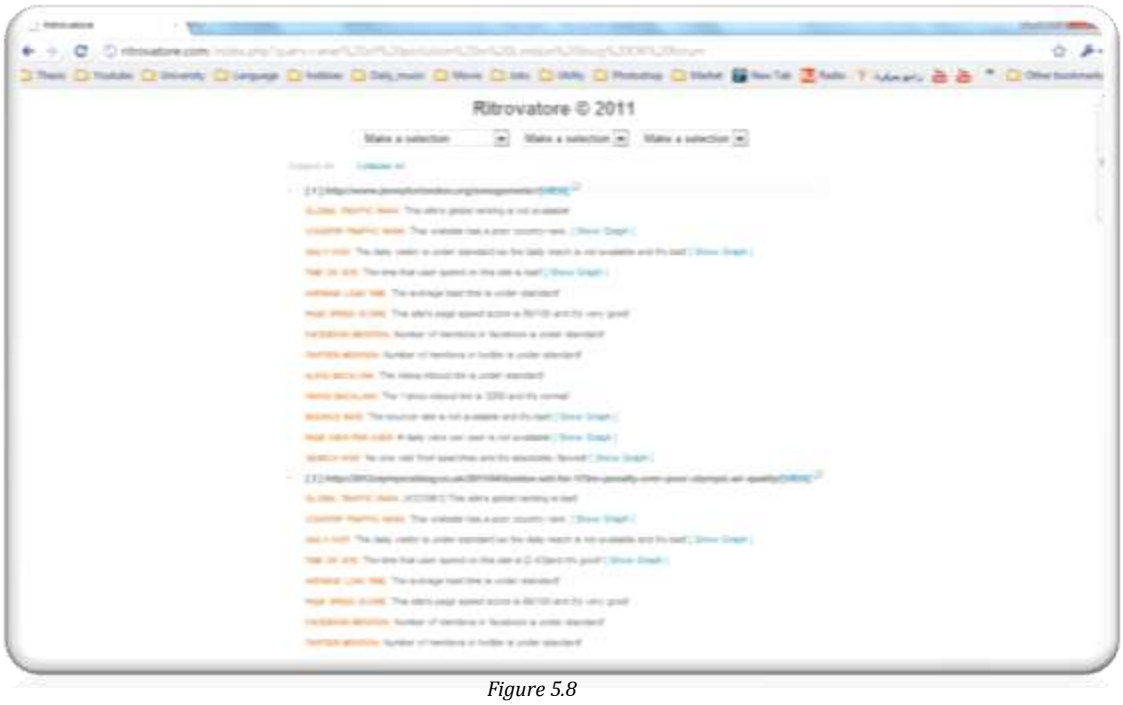


Figure 5.8

Some of the metric according to their provider will give you the blue link at the end of each line name as" VIEW GRAPH". If you click over this button you will see the daily graph related to that metric.

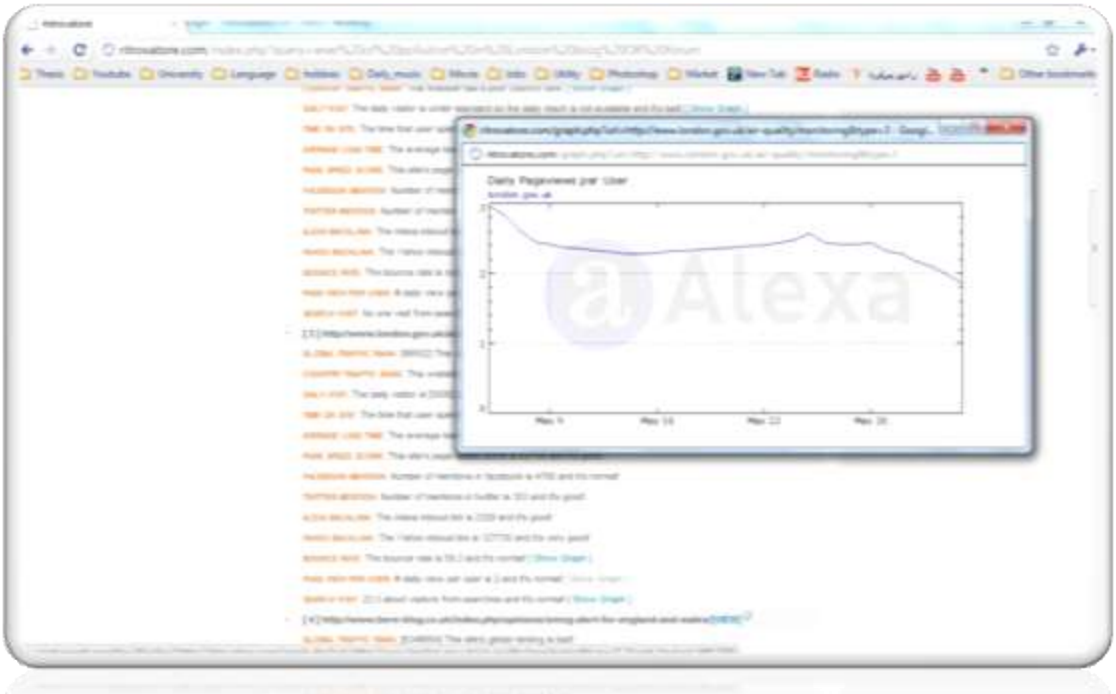


Figure 5.9

2) *Web-reputation based rank:*

- *The second list in the middle of the page is re-ranked order of eight first result of the Google based on our web reputation system. All of the ability that discuss above are executive for this list as well.*

You can easily browsing information and look at the different calculated value base on their metric and our final re-rank order.

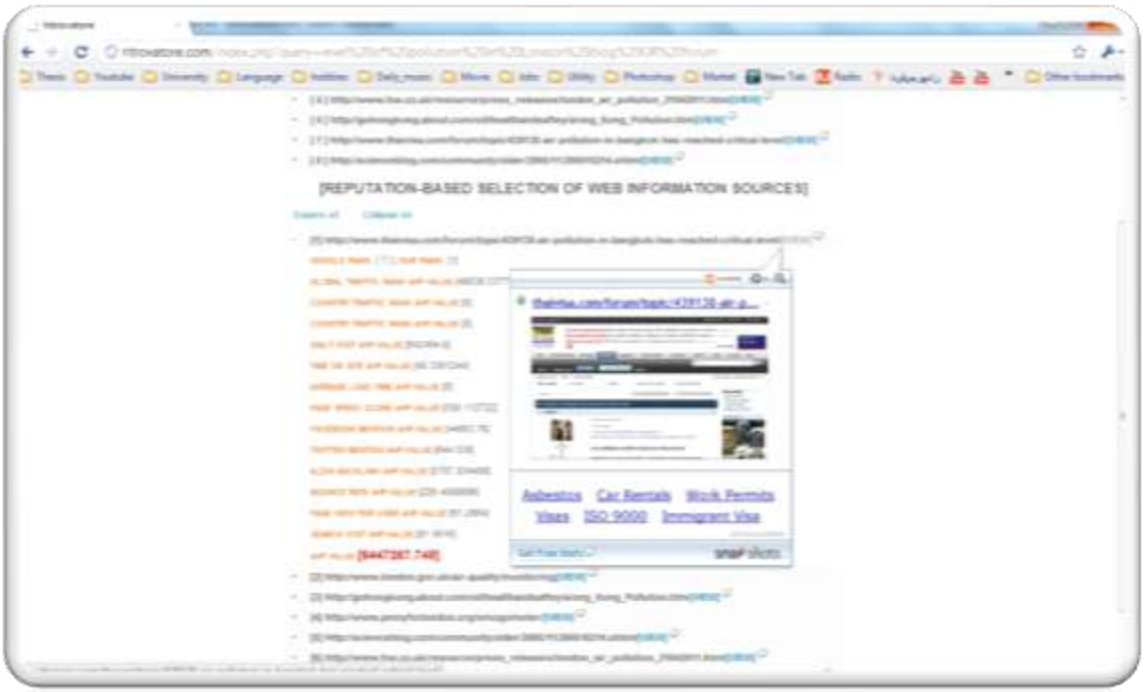


Figure 5.10

REFERENCES

1. Jiang, S., Zilles, S., Holte, R., 2008. Empirical Analysis of the Rank Distribution of Relevant Documents in Web Search.

2. Donato Barbagallo, Cinzia Cappiello, Chiara Francalanci, Maristella Matera. Reputation-Based Self-Service Environments.

3. Donato Barbagallo, Cinzia Cappiello, Chiara Francalanci, Maristella Matera. Semantic sentiment analyses based on the reputation of Web information sources.

4. Chen, K., Zhang, Y., Zheng, Z., Zha, H., Sun, G., 2008. Adapting ranking functions to user preference. Data Engineering Workshop, ICDEW,

5. Gupta, S., Jindal, A., 2008. Contrast of link based web ranking techniques

6. Danette McGilvray. Ten Steps to Quality Data and Trusted Information™; published by Morgan Kaufmann Publishers 2008.

7. Alan R. Tupek, Chair, 2006.Definition of Data Quality.

8. Sean A. Golliher, 2008. Search Engine Ranking Variables and Algorithms

9. Webopedia.com

10. Searchenginewatch.com

11. www.alexa.com

12. www.wikipedia.com

13. www.alvit.de

14. www.reference.com

15. people.revoledu.com

16. www.searchengineoptimizationjournal.com

17. www.executionmih.com

18. www.answers.com

19. www.gfkamerica.com

20. www.doshdosh.com

21. www.mediacollege.com

22. www.blogussion.com

23. www.webconfs.com

24. www.squidoo.com

25. www.thewindowsclub.com

26. www.sitepronews.com

27. http://www.articlesbase.com/business-articles/the-seriousness-of-managing-
28. http://www.submitawebsite.com/services/search-engine-reputation-management.html
29. http://webreputationmanagement.info/
30. http://www.buildingreputation.com/
31. http://searchenginewatch.com/article/2064539/How-Search-Engines-Rank-Web-Pages
32. http://www.scribd.com/doc/52096691/Web-reputation
33. http://www.mikes-marketing-tools.com/ranking-reports/
34. http://www.encyclo.co.uk/define/Kendall%20tau%20distance (kt distance)
35. http://stackoverflow.com/questions/728261/is-it-possible-to-preload-page-contents-with-ajax-jquery-technique
36. http://moofx.mad4milk.net/#gethelp
37. http://code.google.com/p/seostats/#Alexa_Methods
38. http://analytics.mikesukmanowsky.com/analytics/index.php/2008/07/08/measuring-content-effectiveness/
39. www.kryogenix.org/code/browser/sorttable/
40. www.devshed.com/c/a/PHP/Getting-Data-from-Yahoo-Site-Explorer-Inbound-Links-API-using-PHP/3/
41. www.dbuggr.com/smallwei/php-disqus-api/
42. www.radian6.com/blog/2010/03/10-key-content-performance-metrics-to-track/
43. code.google.com/p/gapi-google-analytics-php-interface/
44. www.javaneverdie.com/seo/alexa-com-global-page-views-number/
45. merabheja.com/calculate-adsense-revenue-with-alexa-rank/